DEVELOPING A DIAGNOSTIC MEASURE OF LINGUISTIC COMPETENCE IN THE ENGLISH LANGUAGE FOR LOWER SECONDARY SCHOOL STUDENTS IN SARAWAK

KHO CHUNG WEI

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

DEVELOPING A DIAGNOSTIC MEASURE OF LINGUISTIC COMPETENCE IN THE ENGLISH LANGUAGE FOR LOWER SECONDARY SCHOOL STUDENTS IN SARAWAK

KHO CHUNG WEI

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF EDUCATION (MEASUREMENT AND EVALUATION)

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: KHO CHUNG WEI

Matric No: PMB160002

Name of Degree: MASTER OF EDUCATION (MEASUREMENT & EVALUATION)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

DEVELOPING A DIAGNOSTIC MEASURE OF LINGUISTIC COMPETENCE IN THE ENGLISH LANGUAGE FOR LOWER SECONDARY SCHOOL STUDENTS IN SARAWAK

Field of Study: MEASUREMENT AND EVALUATION

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 08 AUGUST 2019

Subscribed and solemnly declared before,

Witness's Signature Date: 08 AUGUST 2019

Name:

Designation:

ABSTRACT

English language education in Malaysia is undergoing a reform, where students are expected to graduate secondary schools with the ability to use English in daily situations, including the workplace. However, some students enter secondary schools without even acquiring the basic functional literacy in English. The role of English language teachers especially in the lower secondary is to provide remedial instructions to narrow the language gaps of these students. A useful tool to support the work of these teachers is a user-friendly and efficient diagnostic test that can identify the students' language gaps as accurately as possible. This study was an attempt to develop such a test. Using cross-sectional survey design, the test was administered to a representative sample (N=3,086) of lower secondary school students in the southern zone of Sarawak. Data analyses indicate that the item response data best fit the between-item multidimensionality Rasch model, suggesting that the diagnostic English language test measures six related unidimensional latent variables. These variables were positively correlated to each other as expected from a multidimensional test of the same construct. However, not all the items and cases fit the model. Out of 90 items and 3,086 cases, the responses of three items did not fit the Rasch model when the test-takers' abilities were not targeted by the items while 5.83% of the cases underfit the model. The misfits occur most probably due to guessing. There were also 21 problematic items that could not discriminate test-takers with low ability from those with high ability. Moreover, it was found that there were gaps in the item distribution across the range of test-takers' abilities for five of the six dimensions although the overall item difficulties were normally distributed. In terms of differential item functioning, one of the items was found to have moderate to large DIF across grade levels and age cohorts while another item had slight to moderate gender DIF. The DIF

that exist across ethnic groups, native language clusters and geographical areas were considered negligible. The study has demonstrated that the diagnostic test has promising potential, but much work still needs to be done.

Keywords: diagnostic test, English language, Rasch model, multidimensional

PEMBINAAN UJIAN DIAGNOSTIK BAHASA INGGERIS BAGI MURID MENENGAH RENDAH DI SARAWAK

ABSTRAK

Pendidikan bahasa Inggeris di Malaysia sedang mengalami pembaharuan, di mana pelajar dijangka menamatkan pengajian sekolah menengah dengan keupayaan untuk menggunakan bahasa Inggeris dalam situasi harian, termasuk tempat bekerja. Walau bagaimanapun, sesetengah pelajar memasuki sekolah menengah tanpa memperoleh literasi asas dalam bahasa Inggeris. Peranan guru bahasa Inggeris terutamanya di sekolah menengah rendah adalah untuk menjalankan program pemulihan untuk merapatkan jurang bahasa pelajar. Satu alat yang berguna untuk menyokong program pemulihan tersebut adalah ujian diagnostik yang mesra pengguna dan berkemampuan dalam mengenal pasti kelemahan bahasa pelajar. Kajian ini adalah satu usaha untuk membina ujian tersebut. Menggunakan reka bentuk tinjauan keratan rentas, kajian ini telah dijalankan ke atas sampel (N=3,086) representatif pelajar sekolah menengah rendah di zon Selatan di Sarawak. Analisis menunjukkan bahawa data respons item paling sesuai dengan model Rasch multidimensi. Ini bermakna ujian diagnostik tersebut mengukur enam pembolehubah laten unidimensi yang berkolerasi positif. Walau bagaimanapun, bukan semua item dan individu sesuai dengan model. Daripada 90 item dan 3,086 responden, respons untuk tiga item tidak sesuai dengan model Rasch apabila kemampuan individu tidak disasarkan oleh item tersebut manakala respons untuk 5.83% responden tidak menepati model. Ketidaktepatan antara respons dengan model berlaku mungkin disebabkan oleh tekaan. Terdapat juga 21 item yang bermasalah kerana tidak dapat membezakan pelajar yang berkeupayaan rendah daripada yang berkeupayaan tinggi. Tambahan pula, didapati bahawa terdapat

jurang dalam taburan item bagi lima daripada enam dimensi walaupun kesukaran item keseluruhannya menepati taburan normal. Dari segi keberfungsian item differensial (DIF), satu item didapati mempunyai DIF yang sederhana besar dari segi gred dan umur, manakala satu item lagi mempunyai DIF jantina yang sederhana. DIF yang wujud antara etnik, bahasa ibunda dan kawasan geografi boleh diabaikan. Kajian ini telah menunjukkan bahawa ujian diagnostik tersebut mempunyai potensi, tetapi masih perlu diperbaiki.

Kata Kunci: ujian diagnostik, bahasa Inggeris, model Rasch, multidimensi

ACKNOWLEDGEMENT

Firstly, I would like to thank my academic supervisor, Dr Shahrir Jamaluddin, and the readers of my dissertation, Dr Bambang Sumintono, and Dr Mohd Rashid Mohd Saad. Their comments and suggestions have guided me in the writing of this dissertation. Without them, completing this dissertation would have been an impossible mission.

Special thanks to Prof. Trevor G. Bond, Dr André A. Rupp, Dr Tey Nai Peng, Prof. Dr Chua Yan Piaw, Dr Loh Sau Cheong, Dr Lawrence Aeria, and Dr Mohd Zali Mohd Nor, who have advised me on various aspects of this study. Their professional advice has guided me in conducting this study especially in the research methodology, sampling technique, and data analysis. I am also indebted to Dr Boon Pong Ying, who is willing to proof-read my dissertation despite her busy schedule. Any error in this dissertation is a mistake of mine, and under no circumstances, shall they be held liable.

I would like to express my appreciation to my employer, the Ministry of Education Malaysia, for granting me a study leave and awarding me the scholarship to pursue my studies in Master of Education. Without the Ministry's financial support, I would not have the opportunity to conduct this study.

This study would not have been possible without permission from the Educational Planning and Research Division, Sarawak State Education Department, and the school administrators. I am also much obliged to the experts for judging the test items, to the school teachers for administering and invigilating the test, and to the students for responding to the test. Their cooperation is crucial to the success of this study.

Finally, I express my sincere gratitude to my family members and friends who accompany me throughout my academic journey either directly or indirectly. Their words of encouragement and moral support have motivated me to complete this dissertation despite the obstacles faced. This dissertation is dedicated to you.

TABLE OF CONTENTS

CONTENTS **Page**

Chapter 3: Methodology

Chapter 4: Findings

Chapter 5: Discussion and Conclusion

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

LIST OF APPENDICES

Page

CHAPTER 1

INTRODUCTION

In the middle of difficulty lies opportunity (Einstein, as cited in Wheeler, 1979).

This opening chapter discusses the driving force behind the ambitious project of developing a diagnostic English language test for lower secondary school students in Sarawak. As test development is a long-term commitment, the focus of the current study is on the early phase of this 'ambitious' project. The chapter begins with an overview of the issues in English language education before highlighting the needs for the development of a localised diagnostic English language test. The challenges in creating a diagnostic measure of language ability are briefly outlined, and basic details pertinent to the study are described. These provide the foundation to Chapter 2, which further expounds on the 'challenges' in diagnostic language test development; and Chapter 3, which would elaborate on how the study is conducted.

1.1 Background of the Study

English language was once an official language of the Federation of Malaya and, later, Malaysia; however, this ceased to be the case when the National Education Policy was implemented in 1967 (Asmah, 1982). When the Malay language took over the role as the medium of instruction in national schools, English is relegated to the status of a second language and taught as a subject. There was a brief period between 2003 and 2012 when English was used as a medium of instruction, alongside Malay, for subjects in the fields of science and mathematics. When this programme was phased out, the Ministry of Education Malaysia introduced a new policy which aims

to uphold the Malay language and, at the same time, strengthen the English language (Ministry of Education Malaysia, 2010).

Since then, it has become the aspiration of the Ministry of Education Malaysia (2012) to develop bilingual proficiency in every student. The latest initiative to this end is the English Language Education Reform Roadmap (English Language Standards and Quality Council, 2015). The Roadmap is a decade-long plan to implement quality English language education of an international standard. At the heart of the reform is the Common European Framework of Reference (CEFR), a common basis for the elaboration of language syllabi, curricula, examinations, and textbooks, which is internationally accepted (Council of Europe, 2001). Specifically, the Roadmap spells out how to align English language education in Malaysia from pre-school to tertiary education to the standards of the CEFR scale. For example, primary school students are expected to acquire basic functional literacy in English and secondary school students are targeted to be able to use English in daily situations with the potential of using it in the workplace (English Language Standards and Quality Council, 2015).

The interest in English language education is especially profound in the state of Sarawak. Sarawak is the only state in Malaysia that has adopted English as the official language of the state administration alongside the Malay language (Povera, 2015; "Sarawak CM to Continue Adenan's English Policy," 2017). The state government also explicitly promotes greater usage of English in schools so that future generations of Sarawakians will have a good command of the language (Aubrey, 2017a, 2017b). For instance, the state Ministry of Education, Science and Technological Research has issued guidelines on initiatives to improve the level of English language proficiency in Sarawakian schools (Chia, 2017). Among these

initiatives include installing bilingual signboards in all schools and seeking the cooperation of all the 82 state elected representatives to adopt schools in their respective constituencies to improve English language proficiency among students (Aubrey, 2017c; Ten, 2017). The state government's stance on English language is generally welcomed and supported by Sarawakian teachers and students ("We Need to Be Realistic and Practical," 2015).

Unfortunately, despite the government's various policies to enhance English language proficiency, the standard of English language is generally perceived to be deteriorating among younger Malaysians (for example, Wong, 2015). Public criticisms of the declining proficiency of the English language are not baseless, for without a reasonable mastery of the global lingua franca, the younger generation of Malaysians would be at a losing end. Over the past few years, for instance, the main stream media had highlighted cases where Malaysians were unable to pursue further education in reputable varsities overseas (Arukesamy, 2015) nor seek employment in the corporate sector (Syed Jaymal, 2015; Yuen, 2015) due to their poor command of English. In 2015, the news of a thousand medical graduates quitting the profession due to a lack of English language skills sent shockwave throughout the nation (Murali, 2015).

In schools, the mastery of English language among students is equally poor. For instance, Alhadjri (2017) reported that some secondary school students are still illiterate in English although English language is a compulsory subject since primary school. Such anecdotal evidence is reflected in the results of public examinations. For example, it was reported that 22.6% of Year 6 students who sat for the 2016 Primary School Achievement Test (*Ujian Pencapaian Sekolah Rendah*; UPSR) failed to achieve the minimum passing grade for English (Lembaga Peperiksaan, 2016); while only 51.9% of the 2015 Malaysian Education Certificate (*Sijil Pelajaran Malaysia*;

SPM) candidates were eligible for the GCE O Level certificate for the English language (Lembaga Peperiksaan, 2015). The results of these nationwide examinations seem to suggest that, despite years of studying English as a subject in schools, a substantial number of students have failed to master basic English language skills. Similar findings were obtained in a study on English language teaching and learning conducted by Cambridge English (2013) under the commission of the Ministry of Education Malaysia. Specifically, the study found that 87% of Year 6 students, 69% of Form 3 students and 55% of Form 5 students were below the level of independent users of the English language.

On an optimistic note, however, Malaysia is consistently categorised as a country with high English language proficiency from the year 2011 to 2016 according to the EF English Proficiency Index (Education First, 2011, 2012, 2013, 2014, 2015, 2016). In fact, in 2016, Malaysia ranked second out of 19 Asian countries, and 12th out of 72 countries worldwide in terms of English language proficiency (Education First, 2016). This appears to be consistent with the results of the International English Language Testing System, where Malaysia's overall mean band scores was 6.8 for Academic track and 7.0 for General Training track, on a 9-band scale system (IELTS, 2015). It is important to note that the samples of test-takers for the EF English Proficiency Index and the IELTS are self-selected, and thus, biased towards those who are interested in learning the language or those who need to prove their proficiency in the language. In other words, the apparently optimistic results may not be representative of the average Malaysians; hence, appears to contradict the public opinion that the English language standard is declining in Malaysia.

1.2 Rationale of the Study

For many Malaysians, English is a foreign language, notably among those in the rural areas and those who use their own ethnic language for daily communication (Yamaguchi & Deterding, 2016). For Sarawakians, the situation is very much the same as in the general Malaysian population – if not worse, considering that vast areas in Sarawak are rural areas where children have limited exposure to English language in their daily life (Riget & Wang, 2016). For them, English lessons in schools are the only contact they have with the language. This means that their English language teachers must shoulder the heavy responsibilities and moral obligations of improving their proficiency in the language.

For secondary school English language teachers, the colossal task of enhancing students' English language proficiency is made more difficult by the fact that not all students who enter secondary schools have acquired the necessary basic functional literacy in English. In fact, Cambridge English (2013) found that 32% of Year 6 students were operating below the lowest level of generative language use; in other words, they cannot even engage in simple interactions on very familiar topics using English when they entered secondary schools. However, these students are expected to graduate from secondary schools with the ability to use English in daily situations with the potential of using it in the workplace (English Language Standards and Quality Council, 2015). Thus, secondary school English language teachers have no choice but to provide remedial teaching in the shortest time possible so that teaching at the next level can proceed. This implies that the remedial teaching needs to be done in the most effective and efficient way, and this means tailoring teaching to help students overcome their weaknesses. One of the tools that can potentially pinpoint the source of students' weaknesses is a diagnostic language assessment (Lee, 2015).

To date, there are very few English language tests that are purely diagnostic in nature (Alderson, Brunfaut, & Harding, 2015; Alderson, Clapham, & Wall, 1995; Hughes, 2003), and even more scarce in the local context. The few well-known diagnostic English language tests are the Diagnostic Language Assessment System (DIALANG), the Diagnostic English Language Assessment (DELA), the Diagnostic English Language Tracking Assessment (DELTA) and the Diagnostic English Language Needs Assessment (DELNA). These tests aim to diagnose English language needs of young adults either for a general purpose or for academic language needs (Alderson, 2005; Language Testing Research Centre, 2009; Lockwood, 2013; The University of Auckland, 2016); and none of them are designed specifically for the Malaysian or the Sarawakian context.

Since language learning is a social-psychological process within a wider sociocultural context (Arabski & Wojtaszek, 2011) and language assessment is very much related to language learning especially in diagnostic testing, it follows that language testing cannot occur in isolation from the wider sociocultural context. It is then expected that the diagnostic language tests designed for young adults of different countries would contain culture-related elements that may not be suitable for the lower secondary school students in Sarawak. This implies that using existing diagnostic English language tests to identify language gaps of Sarawakian lower secondary school students would not be appropriate. The scarcity of localised diagnostic English language tests provides the impetus for the current study.

The lack of readily available localised diagnostic English language test implies that if secondary school English language teachers were to utilise diagnostic test as a tool to plan their remedial teaching, they will need to prepare their own diagnostic test. However, preparing a diagnostic English language test seems to be beyond the means

of most teachers in the light of a recent survey on Malaysian lower secondary school English language teachers' assessment practices (Ch'ng & Rethinasamy, 2013). Specifically, the findings of the survey revealed that the majority relies on commercial reference books and past-year test papers as their resources in test preparation. The study also concludes that English language teachers apparently lack theoretical understanding of good assessment practices. As such, it is unreasonable to expect lower secondary school English language teachers to develop their own diagnostic tests that enable them to identify their students' areas of weaknesses in a consistent manner.

Given the urgent need for a localised diagnostic English language test, it is appropriate and timely that a study be conducted to develop a diagnostic English language test for lower secondary school students in Sarawak. The current study is an attempt to assist English language teachers who need to make inferences about their students' strengths and weaknesses in language ability so that they can tailor their teaching to the students' language gaps accordingly. As test development is a long-term commitment, this may appear to be an 'ambitious' attempt; hence, the current study concentrates only on the early phase of the test development process. It is hoped that the diagnostic English language test being developed in this study will be of help in supporting the Sarawak state government's effort in enhancing English language proficiency among Sarawakian students.

1.3 Statement of Problem

From the perspective of language testing, a diagnostic language test must be able to identify specific areas of weaknesses in language ability (Bachman & Palmer, 1996); while from the standpoint of psychometricians, a test must be able to

differentiate test-takers with different abilities on the measured trait in a consistent manner (Finch, Immekus, & French, 2016). This means that a diagnostic English language test must be able to reliably measure test-takers' abilities in specific areas of the English language. Therefore, such a test should consist of enough items that can tap into the different language areas so that valid inferences can be made about test-takers' specific areas of weaknesses in language ability. However, developing measures of human ability is not an easy feat; it would be a journey filled with challenges – some of them are outlined below.

Rossiter (2011) wrote that "social science *knowledge* is dependent – entirely – on valid *measurement*" (p. vii); but regrettably, "most measures in the social sciences today lack realism because they do not measure what they are supposed to measure" (p. 2). To make important contribution to scientific progress, researchers in the social sciences should strive to construct measures, and not merely describe the raw data at hand, because "raw data [such as counts] are *not* measures…as [they are] known in the physical sciences" (Bond & Fox, 2015, p. 22). In the physical sciences, measurement is defined as the process of experimentally determining the value of a quantity using carriers of units of measurement or scales, and this process results in a concrete denominated number expressed in sanctioned units of measurement (Fridman, 2012). Because measurement is an experimental estimation of the quantity, there is always an uncertainty or inaccuracy of a measure (Rabinovich, 2005).

Following the above definition of measurement, the raw score of a test cannot be a measure because it does not have any sanctioned unit of measurement. Similarly, transforming the raw test score to percentage correct or a z-score, as per common practice in the reporting of test results, cannot be called measurement (Fridman, 2012). Converting the raw score to a percentile, another common practice in test reporting,

involves rank-ordering and not measuring. Upon careful contemplation, one would eventually come to the realization that raw score, percentage correct, z-score and percentile cannot be measures of human ability as they depend very much on which items are in the test and who the test-takers are, unlike the measure in the physical sciences which chiefly depends on the quantitative value of the property specific to the object of measurement at hand.

Since the goal of the current study is to provide teachers with a test to make valid inferences about their students' weaknesses in language ability, measures of language ability need to be constructed. The use of raw score, percentage correct, z-score or percentile is not satisfactory because the objective of the test is to measure students' language ability, not just describing the students' scores in the test. To construct measures of human ability from observations, various latent trait models can be used – but "of all the models proposed for item calibration and person measurement, the Rasch model is the easiest to understand and the easiest to use" (Wright, 1978, p. 1). Bond and Fox (2015) also recommended the use of Rasch methods in instrument development and theory building as well as at the very beginning of a research project.

The Rasch model however has several requirements of measurement such as unidimensionality, local item independence, equal item discrimination, and the absence of guessing. When the empirical data do not satisfy these requirements, the data would not fit the Rasch model. The problem is that no empirical situation can completely fulfil all the requirements for measurement (Wright & Stone, 1999). This means that there will always be some misfits to the Rasch model; hence, it involves a judgement call from the analyst to evaluate if the extent of misfits can be tolerated for the purpose of the analysis. It is also the task of the analyst to investigate the reasons behind the misfits by checking for empirical indicators.

Another central requirement of measurement is the property of invariance. When measures are invariant, the relative placements of persons on the ability scale are independent of the measuring instruments used so long as the instruments are suitable for the intended purpose (Wu, Tam, & Jen, 2016). This means that items functioning differently in different contexts, such as across different age cohorts, are indicators that measurement invariance has not been achieved. In the context of language testing, Bachman and Palmer (1996) has identified a non-exhaustive list of test-takers' personal characteristics that may influence their test performance; among them are age, gender, native language, and level of education. As the test is intended to be appropriate for lower secondary school students across Sarawak, the test items should function similarly regardless of the students' geographical locations. Therefore, it is important to ensure measurement invariance across these demographic groups.

1.4 Objectives of the Study

The study aims to develop a diagnostic test of English language for lower secondary school students in Sarawak. To ensure that the diagnostic test can reliably measure students' linguistic competence in the English language, the study is set out to achieve the following specific objectives:

1. To assess the dimensionality of the test
2. To investigate the extent to which the item response data fit the Rasch model
3. To estimate the reliability of the data analysed using the Rasch model
4. To determine the extent to which the item difficulty matches the person ability

5. To examine whether the items function differently across the different age, grade, gender, ethnic, native language, and geographical groups

1.5 Research Questions

To address the objectives as outlined above, the study seeks to answer the following research questions:

1. Does the test measure a coherent unidimensional latent variable, several related unidimensional latent variables, or a unidimensional latent variable that can be decomposed into several subdimensions?
2. How well do the students' responses to the test items fit the Rasch model?
3. How reliable are the item placements and the person ordering?
4. How well does the spread of item difficulties match the person ability distribution?
5. Do the items function differently across the different demographic groups, as stated below:
   (a) Age cohorts?
   (b) Grade levels?
   (c) Genders?
   (d) Ethnic groups?
   (e) Native language clusters?
   (f) Geographical areas?

1.6 Significance of the Study

As the focus of the current study is on the early phase of test development, further research is much needed before the diagnostic test of linguistic competence in the English language can be deemed as suitable for state-wide administration. For test consumers who are interested in using the diagnostic test with different populations, for instance with students in other Malaysian states, studies on measurement invariance of the test must be conducted. The items with known parameters can be retained in a pool, and later developed into a computer adaptive test. The reporting of the diagnostic test results, how test consumers make use of the reports, and its effects on instructions are other areas worth investigating especially in the later phase of test development. Further investigation into these areas will help to develop the knowledge and skills of researchers in the fields of psychometrics and language testing at the local level. Therefore, the current study can be perceived as the foundation for test developers who are interested in a localised well-calibrated diagnostic test of English language.

In the current study, the results of the diagnostic test have the potential to provide useful information for English language programme developers. The English language departments at the school level, the school improvement specialist coaches at the district level and the English language education officer at the state level can make use of the test results to design suitable intervention programmes targeted at different groups of students with specific areas of weaknesses. In the later phase of test development, when the results are reported at the level of individual students, English language teachers can make use of the test results to plan lessons that cater to the needs of their students. If the diagnostic profiles are made available to parents, they can provide additional supports to help remedy their children's language gaps. Finally,

students themselves can treat the diagnostic profiles as feedback on what they need to learn to master the English language. These are the long-term significance of the study to test consumers, bearing in mind that the diagnostic English language test is an important tool to help identify specific areas of weaknesses. It is also more economical and sustainable to make use of a locally developed test as compared to using tests that are imported from other countries.

In view of the scarcity of language tests that are purely diagnostic in nature, the current study seeks to fulfil this gap; hence it is of significance to the field of diagnostic language testing. Since the language test development process is much grounded in theories from the fields of linguistics and psychometrics, findings of the study may lead to refinement of relevant models. The application of various techniques throughout the test development process may provide a framework for researchers who are interested in advancing the field of diagnostic testing. Researchers in any of these fields may find the study relevant to their research contexts. In short, the study will be of interest to both practitioners and researchers.

1.7 Limitation of the Study

The current study is limited to the early phase of the test development process, where a new test was designed, and new items were written. How the items would behave during test administration are not known. There is no guarantee that the items would be functioning as expected although multiple precautionary steps are taken in the item writing process. The Rasch model is applied with the intention of prescribing how the items can be retained, modified, combined, or discarded, so that the requirements of measurement can be met when the diagnostic test is used as a measure of linguistic competence.

It is also important to note that the diagnostic test in the current study is limited to measuring only linguistic competence. This means that the test cannot be used as a measure of other components of communicative competence such as sociolinguistic competence and pragmatic competence. It would be ideal for the diagnostic test to be able to measure all the different components of communicative competence; however, it is beyond the scope of the current study to design such a test. Therefore, at this early phase of test development, the focus is on linguistic competence which is the most common and well-operationalised component. Linguistic competence entails various subdomains such as syntax, semantics and phonology. Due to the constraints of the testing situation that is not conducive for listening tasks, the subdomain of phonology is not tested in the current study.

Since the study is at the infancy stage of test development, the diagnostic test needs to undergo further validation studies before it can be used as a standardised test by classroom teachers. Any modification to the items as prescribed by the Rasch analyses need to be further tested in the field. If the test were to be used with a different population such as with lower secondary school students in other Malaysian states, studies of measurement invariance needs to be carried out. Essentially, test development is a never-ending process and studies on the validity of test use need to be conducted frequently.

1.8 Operational Definitions

The diagnostic test is designed to measure linguistic competence, which is one of the main components of language ability. In the literature of language ability, the terms "grammatical knowledge" (Bachman & Palmer, 1996), "grammatical competence" (Canale & Swain, 1980), and "grammatical proficiency" (van Bon, 1992)

are sometimes used interchangeably with the term “linguistic competence” (Rosyidi & Purwati, 2017). To maintain consistency, the term “linguistic competence” is used throughout this manuscript. Linguistic competence is defined as the knowledge of the linguistic code, the ability to recognise the different language features and to manipulate these features to form meaningful words and sentences in accordance with the governing principles or rules (Savignon, 1997). To measure linguistic competence, the test must consist of items that can tap into knowledge, recognition, and/or rule-based manipulation of linguistic codes.

There are six domains of linguistic competence that can be measured in a pencil-and-paper test. The following are the six domains (the terms used in this study are in *italics*) and their definitions:

(a) *Graphology*, also known as orthography, refers to the symbols of which written texts are composed, for example, letter forms, spelling, and punctuations. It is the equivalence of phonology, which is found in spoken texts (Council of Europe, 2001).

(b) *Lexical items*, sometimes called vocabulary, are the words of the language, which include single word forms, polywords such as idioms and phrasal verbs, and collocations (Lewis, 2002).

(c) *Word classes* are grammatical elements such as articles, pronouns, prepositions, and question words (Council of Europe, 2001).

(d) *Morphology* refers to the internal organisation of words including compound words and the use of affixes (Purpura, 2004).

(e) *Syntax* is the organisation of words in a sentence (Lock, 1996).

(f) *Semantics* is the organisation of meanings, for example the relation of words to the general context, and the relations between lexical items (Lock, 1996).

When operationalizing the six domains in the diagnostic test, the national English language syllabi, school textbooks, general grammar books, and linguistics reference books are referred to during the item writing process. The items are written in such a way that they tap into knowledge, recognition, and/or manipulation of rules within the six domains. To ensure that the aggregation of the items can measure linguistic competence without bias towards any domain, equal proportion of items are written for each of the six domains. In other words, the diagnostic test aims to measure the latent variable "linguistic competence" through the six latent domains.

1.9 Summary

The declining standards of English language among young Malaysians and the profound interest to improve the standards of English among students especially by the Sarawak state government have planted the seed of this long-term test development project. Realizing that it is a somewhat 'ambitious' project, the current study concentrates only on the early phase of the test development process. Borrowing the concept of measurement from the physical sciences, the study seeks to develop a measuring instrument of linguistic competence within the framework of Rasch model. However, empirical data will never completely fit the Rasch model; hence, the study is set out to investigate how well the data fit the Rasch model. Findings from the study will inform the later phase of the test development so that future version of the diagnostic English language test is an improvement of the current version. It is only through many cycles of testing out the test that a well-calibrated localised diagnostic test of the English language can materialize.

CHAPTER 2

LITERATURE REVIEW

There is often a belief that 'language testers' have some almost magical procedures and formulae for creating the 'best' test (Bachman & Palmer, 1996, p. 3).

Chapter 1 has justified the needs for a well-calibrated localised diagnostic English language test, highlighted the problems in developing such a test, and defined the latent variable to be measured in the test. In this chapter, language models underpinning the operationalization of the latent variable in the test are discussed first. Next, measurement theories are briefly explained before expounding on the Rasch model. Some understanding of the Rasch model is necessary if the next two chapters were to be fathomable. Various test development models, validation frameworks and models of diagnostic language testing are then laid out to provide the foundation for Chapter 3 and selected past studies on language test development are reviewed to illustrate the status quo of language testing. Chapter 2 concludes with the theoretical framework and the conceptual framework of the study.

2.1 Models of Language

To make inferences about test-takers' language ability based on their responses to test items, it is important for the ability to be defined with certain degree of precision to differentiate it from other characteristics that may affect the responses. When the definition is sufficiently precise, items can be written in such a way that they require test-takers to use their language ability to attain correct or highly-rated responses. Items that can be responded to correctly without using language ability will certainly

confound any attempt at measuring language ability. The question that arises then is: What is language ability?

The definition of language ability lies in the numerous models of language that have been proposed (see Appendix A for diagrams of the different language models). According to McNamara (1996), all language models have three dimensions, namely knowledge, ability for use or performance, and actual language use. The first two dimensions are commonly referred to as communicative competence or communicative language ability (Fulcher & Davidson, 2007). Hymes (1972), generally acknowledged as the father of communicative competence (Bagarić & Djigunović, 2007; Cazden, 1996), has defined communicative competence as the ability to use rules of grammar not only accurately but also appropriately according to the communicative events. Hymes proposed the notion of communicative competence in reaction to Chomsky's (1965) view that linguistic performance is a direct reflection of competence only under the ideal situation where the interlocutors are in a homogenous language community, know their language perfectly well and are not affected by irrelevant factors such as distractions. Hymes opposed to such idealization and argued that Chomsky has omitted the sociocultural aspects from his theory.

Based on Hymes's (1972) notion of communicative competence, Canale and Swain (1980) proposed the first model of communicative competence which comprises of three components: (a) grammatical competence; (b) sociolinguistic competence; and (c) strategic competence. Canale (1983) later expanded on the model by segregating the knowledge of discourse rules from the component of sociolinguistic competence into a separate component called discourse competence; and by including the dimension of actual communication in contrast to communicative competence. Grammatical competence is defined as mastery of the linguistic code, which involves

the ability to recognise and manipulate lexical items and morphological, syntactical, semantical, and phonological rules; while discourse competence is the ability to express and interpret a unified text in different genres using cohesive devices to relate sentence forms and coherence rules to organise meanings (Canale, 1984; Savignon, 1997). Sociolinguistic competence is concerned with appropriateness of language use in terms of meanings and forms within specific social contexts; while strategic competence, comprises both verbal and nonverbal strategies that are used to compensate for breakdowns in communication or to enhance effectiveness of communication such as the use of gestures and paraphrase (Canale & Swain, 1981). Grammatical competence and discourse competence reflect the use of linguistic system while sociolinguistic competence and strategic competence define the functional aspect of communication. Savignon (1997) later presented these four components as an inverted pyramid to suggest a possible relationship between them as overall communicative competence increases.

Building upon the earlier works on communicative competence, Bachman (1990) proposed a model of communicative language ability that redefines the different dimensions and components in Canale's (1983) model. Bachman described language ability as the combination of language knowledge and strategic competence that provides language users with the capacity to implement or execute the competence in appropriate contexts. The model shows that strategic competence, which is a set of metacognitive components, serves the executive function that relates language competence to features of the context where language use takes place and to the language user's topical knowledge of the real world. Bachman and Palmer (1996) later added language user's personal characteristics (e.g., gender and language background)

as a component that relates to strategic competence and made several other modifications to the model.

In Bachman and Palmer's (1996) model, language knowledge is defined as the domain of information in language users' memory that can be used to create and interpret discourse, and this includes two broad areas, namely organisational knowledge, and pragmatic knowledge. Organisational knowledge, which is defined as the ability to control over formal language structures, can be further categorised as grammatical knowledge and textual knowledge; while pragmatic knowledge, which is defined as the ability to create or interpret discourse appropriately, can be divided into functional knowledge and sociolinguistic knowledge. Grammatical knowledge is concerned with how individual sentences are organised, which includes knowledge of vocabulary, syntax, and phonology or graphology. Textual knowledge is involved in producing or comprehending texts that consist of more than one sentence, and this can be divided into knowledge of cohesion and knowledge of rhetorical or conversational organisation. Functional knowledge relates the texts to the language users' intentions, and this includes knowledge of four categories of language functions: ideational (to express or exchange information about ideas, knowledge, or feelings); manipulative (to affect the surrounding context); heuristic (to extend knowledge); and imaginative (to create imaginary world for aesthetic purposes). Sociolinguistic knowledge involves knowledge of the conventions that regulate the proper use of dialects, registers, idiomatic expressions, cultural references, and figures of speech.

Unlike Bachman and Palmer (1996), Canale (1983), and Savignon (1997) who listed discourse competence as one of the components in communicative competence, Celce-Murcia, Dörnyei, and Thurrell (1995) perceived discourse competence as the central component through which all the other competences are manifested. Celce-

Murcia et al. has also included an additional component called *actional competence* to refer to the ability to choose knowledge of language functions, which is analogous to Bachman and Palmer's notion of functional knowledge. Celce-Murcia (2007) later expanded on the model by adding conversational competence and non-verbal competence to actional competence and renaming it as interactional competence. Formulaic competence, which is defined as the fixed and prefabricated chunks of language such as collocations and idioms, is also included in the revised model. The sociocultural, linguistic, and strategic competences are virtually the same as the sociolinguistic, grammatical, and strategic components in Bachman and Palmer's model.

The different components of communicative competence as discussed thus far have been reconceptualised differently by different authors. Purpura (2004), for instance, posited that language knowledge is made up of grammatical knowledge and pragmatic knowledge, where grammatical knowledge refers to grammatical forms at the sentence and discourse levels and their semantic meanings, and pragmatic knowledge refers to the contextual, sociolinguistic, sociocultural, and rhetorical appropriateness, acceptability, or conventionality of the utterances. Meanwhile, Littlewood (2011) perceived communicative competence as consisting of linguistic, discourse, sociolinguistic, pragmatic, and sociocultural competences, where pragmatic competence is redefined as the ability to use linguistic knowledge to convey or interpret meanings in real-life situations (including communication breakdowns), and sociocultural competence is redefined as the cultural knowledge that influences exchange of meanings. The notion of communicative competence has become very influential in language classroom and language testing. For example, the Canadian Language Benchmarks adopted the framework of communicative proficiency in their

national standards for describing and measuring second language proficiency, where communicative proficiency is operationalized as comprising of five distinct components (linguistic, textual, functional, sociocultural, and strategic) with linguistic competence as the core (Pawlikowska-Smith, 2002). Similarly, the Common European Framework of Reference is based on the framework of communicative competence which is operationalised as comprising of three components: linguistic, sociolinguistic, and pragmatic competences, where pragmatic competence can be subdivided into discourse and functional (Council of Europe, 2001, 2011).

It is important to note that the models of communicative competence are different from the skill-and-element model of language proficiency propounded by Carroll (1961) and Lado (1961). In the skill-and-element model, language ability is decomposed into several components of language knowledge employed in the four macro-skills of listening, speaking, reading, and writing. This means that a test of language skills would include communicative competence but not vice versa (Bachman & Palmer, 1984). Bachman and Palmer (1996) argued that it is not useful to think in terms of skills because widely divergent language tasks can be classified together under a single skill and it fails to take into consideration that language use happens in specific situated language tasks. This suggests that a useful diagnostic English language test should not be based on the four macro-skills. Instead, the diagnostic test should be designed based on the components of communicative competence which are needed when carrying out different language tasks.

2.2 Measurement Theories

In the physical sciences, measurement is defined as the process of experimentally determining the value of a quantity using carriers of units of measurement or scales called measuring instruments, and the result of a measurement is the product of a number and a sanctioned unit adapted for the quantity of interest (Fridman, 2012; Rabinovich, 2005, 2013). An example of measurement in the physical sciences is the process of weighing a luggage bag using a portable electronic scale, and the result at the end of the weighing process is a number expressed in kilogramme. Unfortunately, measurement in the human sciences were not as straightforward as measurement in the physical sciences. This is because human abilities are latent traits that cannot be directly observed. The measurement of a specific human ability can only be done by observing a sample of behaviours deemed to be typical of the said ability. A common example of measurement in the human sciences is the process of estimating a student's English language ability level through an English language test.

Unlike measurement of physical objects, measurement of human abilities cannot be done on the human abilities themselves but can only be estimated from the observation of sample behaviours. Although measuring human ability is practically different from measuring physical objects, it is important to approximate the standard of physical measurement in the field of human sciences if human sciences were to progress (Bond & Fox, 2015; Wright, 1967). Applying the definition of physical measurement to the field of language testing, the measurement of language ability therefore can be described as an experimental procedure of determining the value of a test-taker's language ability with the help of a language test, and this process should result in a number expressed in a sanctioned unit.

An important feature underscored in the definition of measurement is that measurement is always as an experimental procedure. This is because the result of measurement is never absolutely accurate due to the unavoidable imperfection of the measuring instrument known as measurement error. For instance, the weight of the luggage bag reported by the portable electronic scale would not be the same as the weight reported by the weighing machine at the airport check-in counter even though it is the same luggage bag. Similarly, the measurement of a student's language ability would never be the same across different tests. The difference is due to measurement errors such as the precision of the measuring instruments. This fundamental relationship between the observed value and the true value in measurement is expressed in the classical test theory.

Classical test theory is one of the oldest measurement theories which originated with Spearman's (1904, 1910) work on measurement error. The theory states that the observed score of an individual in a test is the function of his or her true ability that is being measured and a set of other factors that are random in nature. For example, the observed score of a test-taker in a language test is a function of his true language ability and other factors such as his health condition during the test and the ambiguity of the test items. This relationship is expressed as a linear equation (Meyer, 2014):

$$X_{vt} = T_{vt} + E_{vt} \quad (1)$$

where X_{vt} = the observed score of person *v* on test *t*; T_{vt} = the true score of person *v* on test *t*; and E_{vt} = the measurement error of person *v* on test *t*.

Based on Equation 1, a person's true ability can be inferred from the observed score. In a multiple-choice language test, for instance, a test-taker's language ability can be inferred from the score in the test, which is typically the number of items that he or she has responded to correctly. This assumes that the test-taker needs to use his

or her language ability when answering each item. Therefore, using the observed scores, it is possible to obtain various characteristics of the test items such as item difficulty and item discrimination. Within the framework of classical test theory, item difficulty is defined as the proportion of examinees who responded to an item correctly while item discrimination is given by the point-biserial correlation between the individual item responses and the total test scores (Shultz, Whitney, & Zickar, 2014). An item with more correct responses is considered easier than an item with less correct responses; hence the higher the item difficulty index, the easier the item is. As for item discrimination, a positive and strong point-biserial correlation is ideal as it indicates that test-takers who respond to the given item correctly score higher in the test than those who answer the item incorrectly. Shultz, Whitney, and Zickar (2014) recommended a point-biserial of at least .20. For an item with multiple-choice responses, the point-biserial correlation should be positive for the correct answer and negative for each of its distractors (Millman & Green, 1989).

Theoretically, if a test is administered to a person for countless times in such a way that the resulting scores are statistically independent and identically distributed and if there are no underlying changes to the construct between the test administration, the mean of all the scores is the true score of the person on the test (Raykov & Marcoulides, 2011). In other words, true score is the expectation of the distribution of all possible statistically independent and identically distributed scores obtained by a person on a given test, and this can be expressed as:

$$T_{vt} = \bar{X}_{vt} \tag{2}$$

where $T_{vt} =$ the true score of person *v* on test *t*; and $\bar{X}_{vt} =$ the mean of all theoretically possible observed scores of person *v* on test *t*.

The difference between the observed score of a person on a given test and the theoretical true score is defined as the measurement error, as can be easily rearranged from Equation 1. From Equation 2, given Equation 1, it can be mathematically proven that measurement error has a mean of zero. This leads to the conclusion that random measurement error is completely uncorrelated with the true score. This implies that there is no covariance between the true score and error, hence the composite variance in the observed score can be expressed as:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \quad (3)$$

where $\sigma_X^2 =$ the variance in the observed score; $\sigma_T^2 =$ the variance in the true score; and $\sigma_E^2 =$ the variance in the measurement error.

Equation 3 is central to the notion of test reliability (Finch et al., 2016). Test reliability is conceptualised as the amount of variance in the observed score that is explained by the true score. It can be mathematically defined as the ratio of the variance in the true score to the variance in the observed score as given in the following equation (Yen & Allen, 1979):

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \text{ or } \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \quad (4)$$

where $\rho_{XT}^2 =$ the reliability coefficient given as the squared correlation between X and T.

The ratio of true variance to observed variance, as shown in Equation 4, can also be interpreted as the reliability coefficient alpha (Cronbach & Azuma, 1962). From Equation 4, it is obvious that small error variance leads to high reliability coefficient. Theoretically, if the error variance approaches zero, the reliability coefficient is one. If the observed score variance cannot be explained by the true score, the reliability coefficient is zero. Therefore, the values for reliability coefficient range from zero to one, with values closer to one being interpreted as better (Meyer, 2014).

In other words, reliability coefficient close to one indicates that a large portion of the variance in the observed score is due to true score, hence reproducible; and a reliability coefficient close to zero indicates that most of the observed score variance is mainly due to random measurement error, hence cannot be easily reproduced. For dichotomously scored items, the special case of alpha is the Kuder Richardson Formula 20 (Thompson, 2003).

Although classical test theory is simple to understand and apply, it suffers from several limitations. First, observed scores are ordinal data strictly speaking (Stevens, 1946). This means that it is not appropriate to perform parametric statistical analyses on the observed scores. Secondly, the test reliability coefficient and item statistics can change substantially when data were obtained from different groups of test-takers (Thompson & Vacha-Haase, 2003). For example, when a language test is administered to a group of advanced language users, the item difficulty will be higher (i.e., easier) and the item discrimination will be lower (i.e., less able to differentiate test-takers of differing abilities) as compared to when the same test is administered to the general population. The lack of heterogeneity among the advanced language users will result in a lower reliability coefficient than the one obtained from the general population. Conversely, the performance of test-takers is greatly dependent on the items included in the test. Observed test scores tend to be lower if test-takers are administered difficult items, but higher if they are given easy items. This implies that classical test theory produces test and item statistics that are not invariant across groups of test-takers and test-takers' observed scores that are dependent on the items (Meyer, 2014).

In other words, observed scores do not fulfil the fundamental requirements of measurement. They do not mean the same thing under different circumstances unlike measurements in the physical sciences. When a traveller says his bag weighs 7kg, one

does not need to see the weighing scale he used to know how heavy 7kg is. However, when a student says she has a score of 10 in her English spelling test, one cannot infer how competent she is in her spelling. The score of 10 could mean complete mastery of the words in the spelling list if there are only 10 items, or it could mean that she has not mastered most of the words if there are 50 items being tested. Transforming the observed score to a percentage correct, a z-score or a percentile does not resolve the issue either. Obtaining 20% correct in the spelling test does not allow inference to be made about the student's spelling ability if one does not know which words she has spelt correctly. If all the items she has scored correctly are words like "onomatopoeia", obtaining 20% correct is a somewhat remarkable achievement than if all the correct items are words such as "cat". A z-score of 3 and a percentile of 99 only gives an indication of the student's relative standing in comparison with other test-takers, and not her spelling ability. Therefore, it is somewhat obvious that observed scores and their transformed values are not measures, as reiterated countless of times by writers such as Bond and Fox (2015), Fridman (2012), and Wright (1967). However, this does not mean that classical test theory should be disregarded and observed scores are of no value; it just means that if test developers seek to construct measures, they need to look beyond classical test theory and its observed scores.

The failure of classical test theory to produce good quality measures of human ability leads to the development of item response theory. Item response theory is a family of latent trait models which posits that the interactions of a person with test items are the functions of the characteristics of the person and the characteristics of the test items which is expressed as (Reckase, 2009, p. 12):

$$P(U = u|\theta) = f(\theta, \eta, u) \quad (5)$$

where $\theta =$ the characteristics of the person; $\eta =$ the characteristics of the item; $U =$ the score on the item; and $u =$ the possible value for the score.

The relationship between person characteristics, item characteristics, and item score in Equation 5 is most commonly expressed using a logarithmic function, which can be graphically illustrated as an item characteristic curve (ICC), as presented in Figure 2.1. In the ICC, the probability of correct response on a given item is plotted on the y-axis while the underlying latent trait of interest which is referred to as *ability* is plotted on the x-axis. As a test-taker's ability increases, so does his or her probability of responding correctly to the item. Figure 2.1 shows that the ICC increases most steeply in the middle part of the curve and much less on the left and on the right of the ability continuum, before approaching zero probability on the leftmost and one on the rightmost. The ICC therefore is a monotonic function bounded by 0 and 1 on the y-axis with no lower and upper bounds on the x-axis.

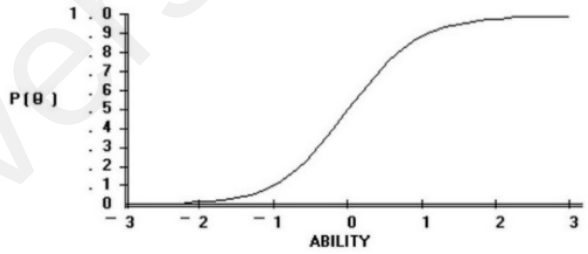
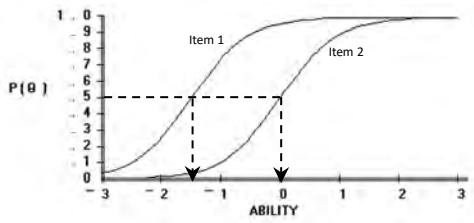


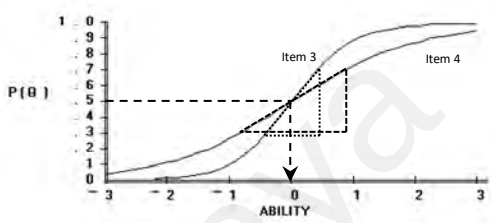
Figure 2.1. Item characteristic curve. Generated using *The Basics of Item Response Theory* (Version 1.0) [Software], by F. B. Baker, 1998. Retrieved from http://ericae.net/irt/baker/

Different features of the ICC represent different item parameters as can be seen in Figure 2.2. First, the position of the ICC denotes the item difficulty and is known as the *b* parameter. Easier items are located to the left while more difficult items are located to the right. The item difficulty index is formally defined as the ability at which about 50% of the test-takers are expected to respond to the item correctly. For example,

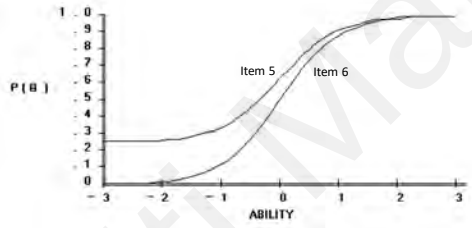
Figure 2.2(a) shows that Item 1 on the left has a difficulty of -1.5 ($b_1 = -1.5$), indicating that test-takers with an ability of -1.5 has a 50% chance of getting the item correct; while Item 2 on the right has a difficulty of 0 ($b_2 = 0$), thus Item 2 is more difficult than Item 1.



(a) Different *b* parameters.



(b) Different *a* parameters.



(c) Different *c* parameters.

Figure 2.2. Comparison of item characteristic curves. Generated using *The Basics of Item Response Theory* (Version 1.0) [Software], by F. B. Baker, 1998. Retrieved from http://ericae.net/irt/baker/

The slope of the ICC at the steepest point on the curve represents the item discrimination which is referred to as the *a* parameter. Figure 2.2(b) shows that the probability of responding correctly to Item 3 changes more rapidly than the probability of responding correctly to Item 4, indicating that Item 3 has a larger *a* parameter than Item 4. This means that Item 3 is better able to discriminate between test-takers with different abilities than Item 4. It is important to note that both Items 3 and 4 are of the same difficulty ($b_3 = b_4 = 0$). However, test-takers with ability below 0 ($\theta < 0$) has a higher probability of responding correctly to Item 4 as compared to Item 3; while test-takers with ability above 0 ($\theta > 0$) has a higher probability of responding correctly

to Item 3 than Item 4. As such, Item 3 will appear to be easier than Item 4 for test-takers with $\theta > 0$, and the reverse is true for test-takers with $\theta < 0$. For test-takers with $\theta = 0$, both Items 3 and 4 appear to be equally difficult as they have the same probability of .50 of responding to the items correctly.

The lower asymptote of the ICC indicates the probability of a test-taker with very low ability responding correctly to the item, and this is referred to as the *c* parameter. Figure 2.2(c) shows that test-takers with very low ability, for instance at $\theta = -3$, have a probability of .25 of responding correctly to Item 5 but a probability of 0 of responding correctly to Item 6. One of the possibilities that test-takers with very low ability can get the item correct is by guessing, hence the *c* parameter is often called the *guessing* or *pseudo-guessing* parameter. According to Lord (1974), the *c* parameter tends to be lower than chance (i.e., .25 in an item with four distractors) in a well-developed test because the distractors would draw low-ability test-takers away from the correct answer. This suggests that if the *c* parameter is high, the item most probably contains non-functioning distractors that even low-ability test-takers can eliminate as possible answers.

The number of parameters used in the logarithmic function to model the ICC determines the name of the models. For example, the three-parameter logistic (3PL) model considers the *a*, *b*, and *c* parameters as shown in Equation 6 (Raykov & Marcoulides, 2011, p. 295):

$$P_{\mathrm{i}}(\theta) = c_{\mathrm{i}} + (1 - c_{\mathrm{i}}) \frac{e^{Da_{\mathrm{i}}(\theta - b_{\mathrm{i}})}}{1 + e^{Da_{\mathrm{i}}(\theta - b_{\mathrm{i}})}} \quad (6)$$

where $P_{\mathrm{i}}(\theta) =$ the probability of correct response on item i; $\theta =$ the person ability; $a_{\mathrm{i}} =$ the item discrimination; $b_{\mathrm{i}} =$ the item difficulty; $c_i =$ the guessing parameter; $D =$ a constant of 1.701 which is a scaling parameter; and e is the base of the natural logarithm.

When the guessing parameter is omitted from the model, or in other words being constrained to 0, the two-parameter logistic (2PL) model is obtained. In the 2PL model, it is assumed that no guessing occurs; and thus, the relationship between the probability of correct response and ability is determined by item difficulty and item discrimination. When all the items are assumed to be equally discriminating, the *a* parameter is fixed to a constant for all the items in the test so that only item difficulty is being taken into account in the ICC. This produces the one-parameter logistic (1PL) model.

There are theoretical reasons and empirical evidence that item response data generally follow the logistic model (Wright, 1977; Thissen & Wainer, 2001); hence the logarithmic function is usually used to model item response data as shown in Equation 6. Apart from the logarithmic function, item response data have also been modelled using the normal ogive function (Samejima, 1977). Many different models have been proposed under the umbrella term "item response theory"; but the easiest to understand in theory and to apply in practice is the Rasch model (Wright, 1978).

As an attempt to meet the needs for individual-centred statistics, Rasch (1960/1980) developed three mathematical models for responses to attainment and intelligence tests. The models specify the distribution function for the possible responses of a person to a given item in a certain test, and the distribution function depends upon a parameter describing the person and another parameter characterizing the item. The first model is a model for oral reading tests where the number of errors made in different reading texts are tabulated; the second is a model for reading speed; while the third model is a model for item analysis of ability tests where the responses are dichotomously scored as 1 if correct and 0 if incorrect. The third model, as

reproduced here in Equation 7 (Rasch, 1960/1980, p. 168), came to be known as the Rasch model:

$$\theta_{vi} = \frac{\xi_v}{\xi_v + \delta_i} \tag{7}$$

where $\theta_{vi} =$ the probability that a person *v* gives a correct answer to an item *i*; $\xi_v =$ a parameter referring to the person; and $\delta_i =$ a parameter referring to the item.

The Rasch model in Equation 7 states that θ_{vi}, the probability of observing a given score in person *v* to item *i*, decreases with increasing difficulty δ_i and increases with increasing ability ξ_v. This is in accordance to the generic function of item response in Equation 5. The Rasch model's item parameter δ_i corresponds to the b_{i} parameter in the 1PL, 2PL, and 3PL models. Since there is only one item parameter being considered in the Rasch model, the Rasch model is mathematically equivalent to the 1PL model (DeMars, 2010). Since the parameters ξ and δ are unknown, the best estimate for θ is the percentage of correct response h such that (Rasch, 1960/1980, p. 76):

$$h \approx \frac{\xi}{\xi + \delta} \tag{8}$$

and

$$1 - h \approx \frac{\delta}{\xi + \delta} \tag{9}$$

Therefore,

$$\frac{h}{1-h} \approx \frac{\xi}{\delta} \tag{10}$$

In a "well-chosen set of test problems" (Rasch, 1960/1980, p. 78) where the possible values of the raw score r are distributed somewhat evenly from low to high values and do not all lump together in the middle, the best estimate for a person's ability parameter ξ can be derived from his raw score r only. This means that the application of the Rasch model to item response data can be used to investigate whether the set of test items is well chosen. If the set of test items is well chosen, the values of the ability parameter ξ that give the same raw score r should have a fairly

narrow range. When ξ is considered as an average ability parameter for persons with the same r (denoted by $\xi^{(r)}$), Equations 8 to 10 are approximately valid. Thereby, Equation 10 can be rewritten as:

$$\frac{\xi^{(r)}}{\delta_i} \approx \frac{h_{ri}}{1-h_{ri}} \tag{11}$$

For practical purposes, Equation 11 can be transformed by taking the logarithms of both sides to obtain the following (Rasch, 1960/1980, p. 80):

$$\log\frac{\xi^{(r)}}{\delta_i} \approx \log\frac{h_{ri}}{1-h_{ri}} \tag{12}$$

$$\therefore \log\frac{\xi^{(r)}}{\delta_i} = \log\xi^{(r)} - \log\delta_i \approx \log\frac{h_{ri}}{1-h_{ri}} \tag{13}$$

The rightmost term $\log\frac{h_{ri}}{1-h_{ri}}$ is the natural logarithm of the odds of correct response (Eckes, 2011), also known as "log of odds unit" (Wu et al., 2016, p. 105), usually abbreviated as 'logit'.

Equation 13 implies that the value of $\xi^{(r)}$ and δ_i can be determined along the same logit scale, where the length of a logit has a consistent value (Bond & Fox, 2015). It is important, however, to note that the length of one logit does not have an absolute meaning. Wu et al. (2016) has demonstrated that the logit scale can shrunk or expand in separate Rasch scaling. This suggests that the length of a logit has a consistent value only within the same scaling, which holds a crucial implication when comparisons are made across scaling such as in the equating of different tests. Nevertheless, Rasch scaling is the only process which can maintain units that support addition (Wright & Stone, 1999). Obvious from Equation 13, its mathematical group structure belongs to the general linear group in the form $x' = ax + b$, which is the operation that determines the equality of intervals (Stevens, 1946); hence taking the natural logarithm of the odds of correct response will produce an interval scale. Therefore, measurements made on the logit scale remains invariant across the intended measurement contexts.

Rasch model also follows the principle of separability. Applying laws of probability on the Rasch model as given in Equation 7, Rasch (1960/1980) has mathematically proven that the item parameter can be estimated independently of the person parameter and vice versa, relying only upon the observed marginals. This means that the Rasch model renders it possible to separate person parameters from item parameters in the data analysis. The principle of separability leads to 'specifically objective' statements about person and item parameters (Rasch, 1966). In making specifically objective statements, the comparisons between persons and items must be free from the conditions under which the comparisons are made. Specific objectivity implies that the measurement of person ability is independent of the spread of items in the test used to measure the ability, and item calibration is independent of the distribution of the persons taking the test. As such, Rasch model produces item-free person measures and person-free item measures (Bond & Fox, 2015; Wright & Panchapakesan, 1969). The specific objectivity of Rasch model can be verified by taking the difference between the log odds of two persons (Wu et al., 2016). Therefore, operating simple arithmetic on Equation 13, it can be shown that the difference between the log odds of person 1 and person 2 when encountering item *i* cancels out the item parameter δ_i, i.e. $\left(\log \xi^{(1)} - \log \delta_i\right) - \left(\log \xi^{(2)} - \log \delta_i\right) = \log \xi^{(1)} - \log \xi^{(2)}$, leaving only the difference between the person parameters.

The remarkable properties of the Rasch model have prompted many extensions of the dichotomous model presented in Equation 7. Some well-known examples include the many-facets Rasch model for item response data that incorporates more than two facets such as raters, interviewers, and scoring criteria (Linacre, 1989); and the testlet model for tests with bundles of items sharing a common stimulus (Wang & Wilson, 2005). In an attempt to unify the different extensions of Rasch model into a

common framework, Adams and Wu (2007, p. 59) proposed the mixed coefficients multinomial logit (MCML) model as a generalized form of the Rasch model:

$$f(\mathbf{x};\xi|\theta) = \left\{\sum_{\mathbf{z}\in\Omega} e^{[\mathbf{z}^T(\mathbf{B}\theta+\mathbf{A}\xi)]}\right\}^{-1} e^{[\mathbf{X}^T(\mathbf{B}\theta+\mathbf{A}\xi)]} \quad (14)$$

where $f(\mathbf{x};\xi|\theta) =$ the regression of the response vector on the item and person parameters; $\mathbf{x} =$ a particular instance of a response; $\xi =$ a fixed set of unknown item parameters; $\theta =$ person ability; $\Omega =$ the set of all possible response vectors; $\mathbf{X}^T =$ the response vector; $\mathbf{A} =$ the design matrix; and $\mathbf{B} =$ the scoring matrix.

The MCML model can be specified into a between-item multidimensionality model and a within-item multidimensionality model (Adams, Wilson, & Wang, 1997). In a between-item multidimensionality model, each item calls upon a single dimension, but the collection of all the items is associated with more than one dimension; while in the within-item multidimensionality model, some items are associated with more than one dimension. To reconcile the apparent contradiction between unidimensional and multidimensional models, Brandt (2008) proposed a Rasch subdimension model as a special case of the multidimensional MCML model. Given that a main dimension can be decomposed into several subdimensions, the subdimension model assumes that the items associated with a common subdimension are more strongly related with each other than with items from other subdimensions. The subdimension model allows for unidimensional and multidimensional Rasch scaling to be conducted concurrently, and is defined as follows (Brandt, 2017, p. 8):

$$\text{Log}\left(\frac{p_{vi}}{1-p_{vi}}\right) = d_{k(i)}\left(\theta_v + \gamma_{vk(i)}\right) - \sigma_i \quad (15)$$

where $p_{vi} =$ the probability of person v responding correctly to item i; $\sigma_i =$ the difficulty of item i; $\theta_v =$ the ability of person v on the main dimension; $\gamma_{vk(i)} =$ person v's specific ability for subdimension k with item i associated to subdimension

k; and $d_{k(i)} =$ the translation parameter that relates the different subdimensional scales to a common scale.

All the Rasch models produce linear, additive values on an interval-level measurement scale which remain invariance within its intended use. To take advantage of these features, the Rasch model (and its various extended models) must hold. When proposing the model, Rasch (1964) has made three important assumptions. The first assumption states that, for each situation of person *v* encountering item *i*, there is a corresponding probability of a correct answer. Therefore, the Rasch model is a probabilistic model as opposed to a deterministic model (e.g., Newton's law of universal gravitation) and a stochastic model (e.g., Mendel's law of heredity). Secondly, the model assumes that the situation of person *v* encountering item *i* is the product of two factors: one pertaining to the person, and another to the item. When the situation involves unknown extraneous factors, the model may no longer be specifiable.

The third and equally important assumption is that all the responses are stochastically independent, given the parameters. Person *v*'s responses to item *i* and item *j* are said to be stochastically independent when the probability of correct response to both item *i* and item *j* are the product of the probability of correct response to item *i* and the probability of correct response to item *j*, i.e. $P\big[(a_{vi} = 1) \cap \big(a_{vj} = 1\big)\big] = P(a_{vi} = 1)P\big(a_{vj} = 1\big)$. Similarly, responses to item *i* by person *v* and person *w* are stochastically independent when the probability of both persons responding correctly is given by: $P[(a_{vi} = 1) \cap (a_{wi} = 1)] = P(a_{vi} = 1)P(a_{wi} = 1)$. The probabilities corresponding to any combination of persons and items are unaffected by all other responses. The third assumption follows from the law of total probability, which views the sample space (i.e. the whole set of responses) as a union of mutually exclusive subsets (Wackerly, Mendenhall, & Scheaffer, 2008).

The remarkable properties of the Rasch model exist only if the data fit the Rasch model (Bond & Fox, 2015; Wu et al., 2016). When the item response data violates the assumptions of the Rasch model, the data may no longer fit the model. Following from Assumption 2, for instance, misfits between the data and the model would occur if the encounter between person *v* and item *i* is affected by unknown factors. For example, person *v* may resort to guessing when item *i* is too difficult for him, or he may be careless when responding to item *i*. Person *v* may also use some other special abilities, knowledge or experience unique only to him, and perhaps to a few others, but not common among the general population of test-takers. On a similar note, item *i* may tap into other ability besides θ that is predominantly measured by all the other items. Item *i* may differentiate the test-takers better than or worse than the other items in the test; or perhaps, it is just poorly written. Violation to Assumption 3 would also cause misfit between the data and the Rasch model. For instance, the answer to one item may provide the clue to another, which may occur when the response to one item depends upon the answer to another; or a person's response is influenced by another person's answer, which can occur when test-takers copy each other's answers.

The scenarios mentioned above are common occurrences in educational testing. This leads to the unfortunate conclusion that no empirical data will ever fit the Rasch model perfectly. Hence, when applying the Rasch model in the analysis of item response data, the analyst seeks for a reasonable degree of fit between the data and the model. What is reasonable depends very much on the purpose of the analysis. Linacre (2017), for instance, recommends a more stringent fit for high-stakes testing than for other practical purposes. Fortunately for the analyst, there are many different fit indices and indicators that can be used to evaluate how well the item response data fit the

Rasch model and to investigate the reasons behind the misfits. Some of these fit indices assess the global model fit while others check for specific violations of the model. Among the specific violations that are usually checked during the analysis of fit are dimensionality of the test, aberrant or unexpected response patterns, poorly written items, item discrimination, item local dependence, and failures of invariance across measurement contexts (Bell, 1982; Bond & Fox, 2015; Douglas, 1982; Meyer, 2014; Wright, 1996; Wright & Mead, 1979; Wright, Mead, & Bell, 1980; Wu & Adams, 2007; Wu et al., 2016).

Oftentimes, the Rasch model is criticised for not fitting the empirical data and that there are many other item response models available for fitting item response data (e.g., Goldstein, 1980; Goldstein & Blinkhorn, 1982). Such criticism highlights a subtle but important difference between fitting a model to empirical data and fitting data to a theoretical model. The former takes a descriptive approach in data analysis, where the goal is to describe the data at hand; while the latter operates from a prescriptive stance, where the goal is to construct measures from raw data. The application of Rasch model to item response data is not to merely describe the data, but more importantly, to construct an interval scale so that the variables of interest can be measured. As such, the failure of data fitting the Rasch model is an indication that "the data do not support the construction of measures suitable for stable inference…[and] they don't add up to anything that lies along any one line of inquiry" (Linacre, 1996, p. 512). The Rasch model, therefore, plays an important role as quality control in instrument calibration and serves as "a tool for construct validity" (Bond & Fox, 2015, p. 343). Since the aim of the study is to develop a measure of linguistic competence, the item response data collected should be fitted to the Rasch model.

2.3 Test Development and Validation

Generally, the process of test development begins with a conceptualization of the test and drafting of the items before the test is administered in the field. Based on the data gathered from the test administration, the test is evaluated and subsequently revised. It is important to note that, in practice, the process is not sequential as the decisions made at one stage may lead to revision of previous stages before progressing to the next stage. The iterative nature of the test development is reflected in the models presented by Bachman and Palmer (1996), Fishman and Galguera (2003), and Markus and Borsboom (2013), and reproduced in Appendix B.

Bachman and Palmer (1996) organised test development into three conceptual stages: (a) design; (b) operationalization; and (c) administration. In the design stage, the test developer describes the purpose of the test, the tasks in the targeted domain and the characteristics of the test-takers before defining the construct either based on syllabus or theories and devising a plan for evaluating the qualities of test usefulness and allocation of resources. The operationalization stage involves developing the test blueprint, writing the test tasks, and specifying the scoring method; while the administration stage involves administering the test, collecting feedback, analysing the items, estimating the reliability of test scores, and investigating the validity of test use. At each stage, consideration must be given to the qualities of test usefulness. The different stages are connected using double-headed arrows to emphasize the iterative nature of test development.

Unlike Bachman and Palmer (1996), Fishman and Galguera (2003) visualizes test development as a sequential process which starts off with conceptualization of test components based on literature review and drafting of items. This initial item pool forms the first version of the test which undergoes pilot-testing and administration.

During the pilot test, some test-takers are interviewed for their opinions on the test to alert test developer of potential issues that might arise during the test administration and test use. The test administration will provide item responses for the analyses of item difficulty, inter-item consistency, item discrimination and instrument reliability. Based on the analyses, items are selected, or new items are drafted for the second version of the test. The second version of the test is administered and analysed before items are selected or revised for the third version. Validity evidence related to external criterion is also collected. The process is repeated for the third and subsequent versions of the test, implying that test development is a never-ending process.

The iterative nature of test development is perhaps best illustrated in Markus and Borsboom's (2013) integrative model. According to this model, testing occurs in a larger context (extra-psychometric criteria) which is usually beyond the control of test developer. Within the larger context, the testing goal, construct, and test procedure are specified. In specifying the construct, it is important that test developer refers to existing theories and makes use of expert judgement. The preliminary test is piloted, and data are gathered. The test developer then connects the theoretical construct to the empirical data by deriving and testing hypotheses. This may lead to revision of the test, and the whole process is repeated. When the test works out as expected, the test is ready to be used; and the consequences of test use and the underlying theories of the construct can be evaluated.

Implicit in all the three test development models is the concept of validity. Validity is the most fundamental consideration in test development (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999); yet, it is the most elusive of all concepts in testing as can be seen in the lack of consensus over

its definitions (Newton, 2013). Validity is classically defined as "the degree to which a test or examination measures what it purports to measure" (Ruch, 1924, p. 13) and is often contrasted with reliability which is a problem of "how consistently it measures" (Buckingham et al., 1921, p. 80). Reliability is often perceived as a necessary but insufficient prerequisite to validity.

The conception of validity originated with the needs for a test to predict a criterion performance; but with the increasing use of achievement tests, it shifted to the question of adequate mapping of test items onto the content domain (Taylor, 2013). The former is known as criterion-related validity while the latter is often called content validity (APA, AERA, NCME, 1966). Cronbach and Meehl (1955) proposed the third type of validity, namely construct validity, which views test score as a measure of an underlying construct. These different conceptions lead to a fragmented view of validity as reflected in the APA, AERA, and NCME's (1954) recommendation that a test manual "should indicate clearly what type of validity is referred to" (p. 18). It is now generally recognised that the traditional conception of validity is a misnomer; hence instead of referring to the different types of validity, AERA, APA, and NCME (1985) referred to them as criterion-related evidence, content-related evidence, and construct-related evidence of validity respectively.

Messick (1989) criticised that the traditional conception of validity is fragmented and incomplete, so he redefined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). To exemplify this unitary concept of validity, Messick (1989, 1994) proposed a unified validity framework with two interconnected facets that crossed to obtain a four-

fold progressive matrix which highlights evidence and consequences in test interpretation and test use (see Appendix B).

It is pertinent to note that construct validity appears in every cell of the framework because it is "the integrating force that unifies validity issues into a unitary concept" (Messick, 1994, p. 24). According to Messick, there are six distinct aspects of construct validity, which can serve as standards or criteria for all educational and psychological measurement. The six aspects are content (relevance and representativeness), substantive (theoretical rationales and process models for response consistencies), structural (congruence between scoring procedures and structural relations of the construct domain), generalizability (generalising from the tasks measured to the broad construct domain), external (relationship between test scores and other measures), and consequential (intended and unintended consequences of test score interpretation and test use). Messick's unified view of validity was adopted by AERA, APA, and NCME (1999) which advocated for a unitary concept of validity that can be established by obtaining evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing.

The inclusion of consequences of testing in the framework of validity has drawn much criticism. The primary concern is that burdening test validation with appraisal of consequences would result in confusion which is detrimental to the evaluation of test score meaning and test use (Popham, 1997). It would also create practical challenges for test developers in obtaining evidence on actual consequences of test interpretation and test use since they do not have control over how their tests are used and interpreted (Green, 1998; Reckase, 1998). Furthermore, Shadish, Cook, and Campbell (2002) argued that evaluation of test use exceeds validation because validity provides only one of the evaluation criteria while other criteria such as cost-

effectiveness and social fairness are beyond the scope of validity theory. As such, evaluation of the consequences of test interpretation and test use are beyond test validation; and thus, they should be distinguished from each other (Cizek, 2012; Scriven, 2002). It must be noted that these critics did not object to the evaluation of consequences of test interpretation and test use; but disagreed with its inclusion within the notion of validity. They prefer a relatively narrow conception of validity that restricts validation within the scope of obtaining evidence bearing on measurement, excluding evidence bearing on consequences.

As an attempt to dissolve the confusion regarding the conception of validity, Newton and Shaw (2014) abandoned the term *validity* in their proposal of a neo-Messickian framework for the evaluation of testing policy. In the neo-Messickian framework, Newton and Shaw distinguished between evaluation of technical quality and evaluation of social value for the three mechanisms within a testing policy (see Appendix B). Each mechanism is framed in terms of objectives, namely measurement objectives, primary decision-making objectives, and secondary policy objectives respectively. Evaluation of technical quality for the measurement objectives considers both the theoretical plausibility and practical viability of the testing procedures; while evaluation of technical quality for the decision-making objectives calls into question the relevance of the test attributes to the outcomes of the decision-making process. These two cells are somewhat akin to the restricted notion of validity as advocated by authors such as Cizek (2012), Popham (1997), and Shadish, Cook, and Campbell (2002). Evaluation of technical quality for secondary policy objectives involves investigating the positive impacts as anticipated by the test developers; while evaluation of social values is concerned with the credibility, utility, fairness, and legality of the testing policy including both anticipated effects and unintended

consequences. The final cell called *Overall Judgement* is a synthesis of all the other cells into an integrated evaluative judgement as envisaged by Messick (1989).

In practice, test validation cannot be separated from test development. This is evident in the considerable overlaps of elements between the test development models and the validation frameworks. The overlaps make it possible for test developers to maximize the validity arguments from the outset before collecting evidence to support or disprove these arguments. In the context of diagnostic language testing, the validity arguments pertaining to the relevance and utility of the test are of utmost importance. This is because the defining characteristics of a diagnostic language test include its test use. Specifically, a diagnostic language test is defined as an instrument developed primarily to identify test-takers' weaknesses and strengths in the targeted language and provide specific diagnostic feedback that can lead to future treatment or intervention (Alderson et al., 2015; Lee, 2015).

Harding, Alderson, and Brunfaut (2015) and Lee (2015) proposed two highly similar models of diagnostic language testing, as reproduced in Appendix C. Central to both models are the notion of identifying specific subskills that need attention before testing students on the subskills. Teachers can either make use of existing tests or develop their own. In practice, developing diagnostic language tests is beyond the means of most language teachers as they have a limited understanding of assessment fundamentals (Ch'ng & Rethinasamy, 2013; Malone, 2013); hence, most teachers are expected to use existing tests. After administering and scoring the test, the diagnostic evidence collected is used to formulate feedback that is linked to a follow-up plan. This implies the needs for a diagnostic language test to be relevant and useful.

A relevant and useful diagnostic language test must be able to offer information about aspects of the language that students need to develop (Brown & Abeywickrama,

2010). For example, a diagnostic language test may identify gaps in the student's knowledge of vocabulary, syntax, semantics, and so forth. Any language test virtually has some potential for providing diagnostic information (Bachman, 1990). Although the notion of retrofitting existing language tests for diagnostic purposes is theoretically possible, in practice however, it has severe limitations (Fulcher & Davidson, 2009). For instance, a placement test that aims to determine the suitable entry level for a language course may lack the specificity and direct item-by-attribute relationship to extract rich diagnostic information about language gaps (Kim, 2015). This implies that for a language test to be truly diagnostic, it must be designed from the outset with a specific diagnostic intent in mind. Unfortunately, there are very few language tests that are purely diagnostic in nature, perhaps, due to some practical problems and theoretical difficulties (Alderson, 1981; Alderson et al., 2015; Hughes, 2003).

To overcome the inadequate theorisation of diagnostic language testing, Alderson et al. (2015) highlights five principles of diagnostic language testing based on findings from interviews conducted with professionals in the fields of automobile, information technology, medicine, psychology, and education. The principles are: (a) it is the user of the test that diagnoses, not the test itself; (b) the test should be targeted, discrete, user-friendly, and efficient; (c) the diagnostic process should involve diverse stakeholder views, including students' self-assessments; (d) the diagnostic process should ideally be embedded within a system that allows for observation, initial assessment, use of tools, test and expert help, and decision-making; and (e) the test should relate to future intervention that can lead to improvement in students' language competence.

The most salient feature of a diagnostic language test is that it needs to be based on specific areas of language knowledge which have been covered or will be covered

in the near future and underpinned by some detailed theory of language (Alderson, 2005, 2007; Cumming, 2015). Such tests are more likely to focus on specific elements than general language proficiency and are discrete-point (testing one element at a time, item by item) rather than integrative (combining many language elements in the completion of a task). Alderson also postulates that a diagnostic language test is usually low-stake, hence removing test anxiety and other affective barriers that might prevent optimum performance. With new computing technology, diagnostic language tests can be enhanced using computers, for instance through computer adaptive testing, although Cumming warns that it will be difficult to be implemented efficiently on a large-scale basis over time.

2.4 Review of Selected Past Studies

A systematic review of EBSCOhost was conducted in December 2017 using the combined keywords "language AND testing AND development" and "language AND test development". The search was limited to scholarly peer-reviewed journals between 2010 and 2018, arranged in order of relevance. From the abstracts of the first 200 articles, there are 52 articles relevant to the field of language testing. As none of the 52 articles are related to the Malaysian context, an additional search was done on MyJurnal, which collects and indexes all Malaysian refereed and scholarly journals. This resulted in eight relevant articles. Additionally, searches using the keyword "test validation" and limited to articles from 2010 to 2018 were also conducted on three journals with an explicit focus on language testing, namely "Language Testing", "Language Testing in Asia" and "RELC Journal", which yielded 114, 31, and 10 relevant articles respectively. Out of the 215 articles, there are only 33 articles related

to diagnostic language testing. This seems to support Alderson et al.'s (2015) contention that there are very few language tests that are purely diagnostic in nature.

In fact, almost half of the diagnostic tests mentioned in the 33 articles are designed for clinical diagnosis of language impairment. For instance, Letts, Edwards, Schaefer, and Sinka (2014) describes the development of the New Reynell Developmental Language Scales (NRDLS), a standardized test to identify language delay among children, and pinpoint their specific difficulties. The test was later adapted to Mandarin (NRDLS-M) by Lim and Lee (2017). Both tests assessed important aspects of language acquisition in terms of comprehension and production within the subdimensions of vocabulary, grammaticality, morphology, and sentence structure. The NRDLS and NRDLS-M were standardised on samples of children, aged 2 to 7, in England (n=1266) and in Malaysia (n=40) respectively. Both studies however did not use random samples in the standardization process; instead, the samples were recruited from schools, nurseries, and homes via personal contacts.

Similarly, Smyk, Restrepo, Gorin, and Gray (2013) developed the Spanish-English Language Proficiency Scale (SELPS) to assess the oral language skills of bilingual children, aged 4 to 8, within the subdimensions of syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity, using story retell tasks. Other similar tests include the French version of the Test for Reception of Grammar (F-TROG; Facon & Magis, 2016), which assesses various linguistic constructions such as embedded sentences; and the Katzenberger Hebrew Language Assessment for Preschool Children (KHLA; Katzenberger & Meilijson, 2014), which assesses auditory processing, lexicon, grammar, phonological awareness, semantic categorization, and narration of picture series. The NRDLS, NRDLS-M, SELPS, F-TROG, and KHLA appear to support Alderson's (2005, 2007) observation that

diagnostic tests are more likely to focus on specific elements than general language proficiency.

In contrast, some diagnostic language tests focus on the four macro-skills. For example, Kostelecká, Kostelecký, Vodičková, and Jančařík (2015) developed a special diagnostic instrument to test students' mastery of the Czech language in reading, writing, listening, and speaking. On the other hand, some diagnostic tests focus only on one of the four skills, for example the Reading Evaluation and Decoding Systems (READS; Abdul Rashid, Lin, & Shaik Abdul Malik, 2012), the Diagnostic College English Speaking Test (DCEST; Zhao, 2013), the Direkt Profil (Granfeldt & Ågren, 2014), the Criterion®, and the Intelligent Academic Discourse Evaluator (IADE; Chapelle, Cotos, & Lee, 2015). READS was developed to identify the English reading ability of secondary school students in Malaysia using multiple-choice items, and reported the test results in performance bands with specific descriptors of what the students can and cannot do; while DCEST was designed as a face-to-face interview test to identify strengths and weaknesses in the speaking ability of university students in China by providing test-takers with a profile test score and a feedback report. Meanwhile, the Direkt Profil, the Criterion®, and the IADE are automated writing evaluation systems which analyse students' essays and provide detailed error feedback on linguistic structures. Tests such as READS, DCEST, the Direkt Profil, the Criterion®, and the IADE can be used by students to bridge their language gaps; and by teachers to improve their students' weaknesses.

The diagnostic language tests that are developed must be validated either during the test development process or when the test is used in the field. There are various methods of obtaining validity evidence. To obtain construct-related evidence, the underlying latent structures of the tests can be examined using factor analysis (e.g.,

Hoffman, Loeb, Brandel, & Gilliam, 2011; van Steensel, Oostdam, & van Gelderen, 2012; Zhao, 2013), the Rasch model (e.g., Mizumoto, Sasao, & Webb, 2017), and cognitive diagnostic models (e.g., Kim, 2011; Li, Hunter, & Lei, 2016; Yi, 2017). To obtain criterion-related evidence of validity, the data from the tests are correlated with some external criteria that are collected either at the same time as the test administration (concurrent validity; e.g., Fletcher, Hogben, Neilson, Lalara, & Reid, 2015; Letts et al., 2014) or at some other time in the future (predictive validity; e.g., Carson, Boustead, & Gillon, 2014). To provide evidence that the test is functioning in the same manner across different groups of test-takers, some studies investigated differential item functioning (e.g., Facon & Magis, 2016; Li & Suen, 2012). Another piece of evidence that is important in supporting the validity arguments of test use is the reliability of the test scores (e.g., Fletcher et al., 2015; Letts et al., 2014).

However, there is no study that can support all the validity arguments for any test. For instance, in a review of 10 standardized unidimensional tests of vocabulary for children under 18, Bogue, DeThorne, and Schaefer (2014) reported that none of the tests passed all the reliability criteria, and only one test met three out of the four validity criteria. In another review, Friberg (2010), who evaluated nine standardised tests for identification of language impairment among pre-school and school-age children against a list of 11 psychometric criteria, reported that all the tests satisfied eight to ten criteria, and eight of the tests provided evidence of item analysis. However, only two of the tests met the predictive validity criteria, and five met the test-retest reliability criteria. It can be concluded from Bogue et al.'s and Friberg's reviews that no tests can satisfy all validity criteria; hence, the mythical 'best' test never exists.

2.5 Theoretical Framework

To address the challenges of constructing diagnostic measures of language ability as stated in Chapter 1, the current study is primarily grounded in models of test development, validation frameworks, and models of diagnostic language testing. Instead of subscribing to any one model, the common elements found in the different models are synthesized to form the theoretical backbone of the current study. Table 2.1 shows the common elements of the different models of test development and test validation that have been presented previously.

Table 2.1
Summary of Different Models of Test Development and Validation

Models of Test Development and Validation	Elements						
	Design	Operationalization	Administration	Construct validity	Item properties	Usefulness	Consequences
Bachman & Palmer (1996)	✓	✓	✓	✓		✓	
Fishman & Galguera (2003)	✓	✓	✓	✓	✓		
Markus & Borsboom (2013)	✓	✓	✓	✓	✓	✓	✓
Messick (1994)				✓		✓	✓
Newton & Shaw (2014)				✓	✓	✓	✓
Harding et al. (2015)	✓	✓	✓			✓	
Lee (2015)	✓	✓	✓	✓		✓	

Note. ✓ Included in the models.

Test design, operationalization, and administration represent the key stages in test development, and they can be further decomposed into different sub-processes. For instance, test design involves specifying the testing goal, stipulating the test procedures, and defining the construct at a level that is suitable for the use of the test. Implicitly, test usefulness and construct validity are deliberated upon from the outset of test design. The operationalization stage transforms the specifications of test design

into items, and this involves drafting of items, judgement from experts, item modification and item selection. The item writers and experts must bear in mind how the items can tap into the construct, and not violate the assumptions of measurement models. When enough items are written, they are put together into a test and administered in the field. The responses to the items are recorded and analysed to obtain empirical evidence pertaining to the validity of the construct and the properties of the items. Reactions from test users can also be collected to evaluate the usefulness and consequences of the test. The results from the last stage becomes input for the next phase of testing, suggesting that test development is an iterative and never-ending process. The different elements in test development and validation are thus interrelated and at times almost practically inseparable.

In the context of diagnostic language testing, the different stages of test development are informed by language models and measurement theories. In the design stage, the construct must be specified in such a way that users can make inference about the language gaps of test-takers. The construct of language ability, therefore, must be specified at a finer grain in a diagnostic test than in an achievement test that requires only a global indicator of language ability. In the operationalization stage, the items must be written in such a way that they follow the principles of diagnostic language testing, reflect the construct that has been specified, and do not violate the assumptions of the measurement models that are to be used for analysis of item response data. Findings from the data analysis will inform the theory of the construct. The interrelationship between test development, construct theory, and measurement theory forms the theoretical basis for any attempt at test development. Figure 2.3 at the end of this section illustrates such an interrelationship in the context of the current study.

In the current study, the construct theory is grounded in the different models of language ability (see Appendix A). Despite the different conceptualization of language in the different models, there are no major theoretical disagreements among them (Brown, 2014). All the models generally agree that language ability is best understood as a multicomponent construct and that individuals may develop the components differentially. The operationalization of language ability in testing depends on the purpose of the test, the constraints of the testing context, and the inferences to be drawn from the test scores (Purpura, 2008). Since the language test to be developed in the current study is intended to be diagnostic in nature without much constraints from any authoritative agency, the construct of language ability must be defined in accordance with what is most common in the field. Table 2.2 summarizes the different models of language in search of the most common component.

Table 2.2
Summary of Different Models of Language

Models of Language	Components					
	Linguistic	Discourse/ Textual	Sociolinguistic/ Sociocultural	Functional	Strategic	Pragmatic
Canale (1983)	✓	✓	✓		✓	
Savignon (1997)	✓	✓	✓		✓	
Bachman & Palmer (1996)	✓	✓	✓*	✓*	✓	✓
Celce-Murcia (2007)	✓	✓	✓	†	✓	
Purpura (2004)	✓	†	†*			✓
Littlewood (2011)	✓	✓	✓		†*	✓
Pawlikowska-Smith (2002)	✓	✓	✓	✓	✓	
Council of Europe (2001, 2011)	✓	✓*	✓	✓*		✓

Note. ✓ These components are explicitly included in the models.
* These components are also subsumed under the pragmatic component.
† Elements of these components are implicitly included in the models, but not named as such.

As can be seen from Table 2.2, the linguistic component is explicitly included in all the models, suggesting that it is crucial to the construct of language ability. The discourse/textual and sociolinguistic/sociocultural components are explicitly named in all the models, except in Purpura's (2004) where they overlap with the grammatical and pragmatic components respectively. The scaling of items in the sociolinguistic component has also been reported to be problematic in past studies (e.g., Council of Europe, 2001). The strategic component is included in more than half of the models while the functional component appears in less than half of the models. The most problematic component in terms of operationalization perhaps is pragmatic competence where it is defined differently in different models and often subsumes other components (e.g., Bachman and Palmer, 1996; Council of Europe, 2001, 2011) or overlaps with other components (e.g., Littlewood, 2011; Purpura, 2004). Only the linguistic component is the least problematic and the most common component in the literature; therefore, this is the construct to be tested in the current study.

Figure 2.3 illustrates the theoretical framework of the current study as an interplay between stages of test development, theory of linguistic competence, and Rasch model. Linguistic competence is defined as knowledge of the linguistic forms and their meanings (Purpura, 2004). It includes the ability to recognise the different word classes as well as the lexical, morphological, syntactical, and graphological features of a language with their semantics and to manipulate these features to form meaningful words and sentences in accordance with the governing principles or rules (Savignon, 1997). In the current study, there are six subdimensions that made up grammatical knowledge: (a) lexical items; (b) word classes; (c) morphology; (d) syntax; (e) semantics; and (f) orthography. Lexical items refer to the vocabulary of English including fixed expressions such as idioms and collocations and single word

forms that have one or several distinct meanings; while word classes are grammatical elements such as articles, pronouns, prepositions, and question words (Council of Europe, 2001). Morphology is concerned with the internal organisation of words including compound words and the use of affixes while syntax deals with the organisation of words in a sentence (Purpura, 2004). Semantics is the organisation of meanings such as the relation of words to general context and the relations between lexical items; while orthography or graphology is the symbols of which written text are composed such as letter forms, spelling, and punctuations (Council of Europe, 2001). Phonology, which is the spoken equivalent of graphology, is not tested in the current study due to the constraints of the testing context that is not conducive for listening tasks.

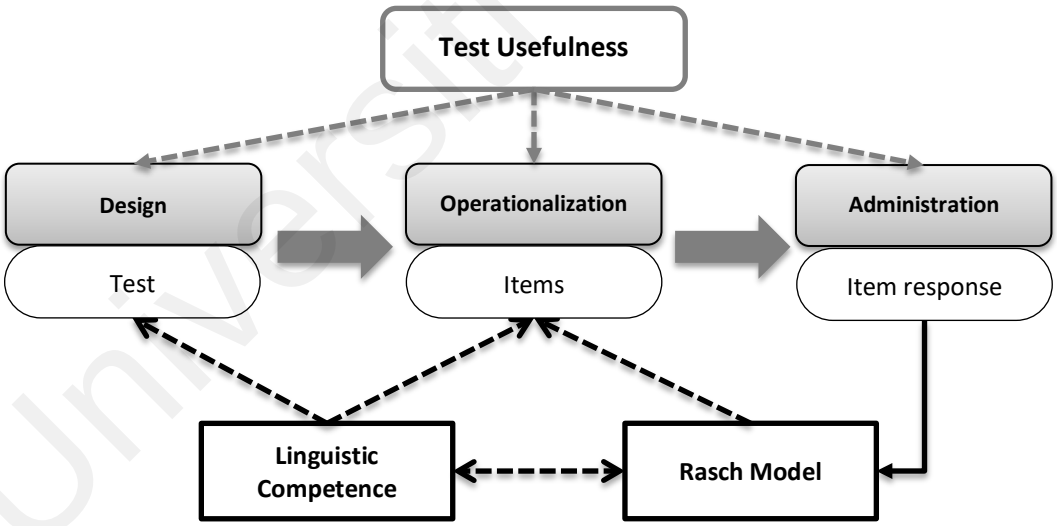


Figure 2.3. Theoretical framework of the study.

Since diagnostic language tests need to be based on specific areas of language knowledge which have been covered or will be covered in the immediate future (Alderson, 2005, 2007; Cumming, 2015), contents from the relevant syllabuses will be mapped into the six subdimensions in the test design stage. The mapping between

the syllabus contents and the subdimensions of linguistic competence forms the test specifications. In the operationalization stage, items are written in areas of topical knowledge familiar to most of the intended test-takers. Students' personal characteristics such as age, level of education, gender, and native language, which may influence test performance, are also taken into consideration in the operationalization stage. As diagnostic language tests are specific and discrete-point (Alderson, 2007; Alderson et al., 2015), each item is written so that it can be associated with only a single subdimension. The most suitable item response format in this case is the multiple-choice format which permits scoring to be done quickly prior to making diagnostic and intervention decisions (Osterlind, 2002). Other item response formats such as essay writing or answering open-ended questions are not suitable for diagnostic language tests as they take more time to score, introduce rater-related measurement errors, and are not discrete-point because they may tap into more than one subdimension.

In the administration stage, the item response data are analysed using the Rasch model because the goal is to produce measures of linguistic competence. Aryadoust (2009) has also demonstrated that Rasch analysis supports validity arguments. Since the test design assumes that linguistic competence can be decomposed into six subdimensions, the measurement model of choice is the Rasch subdimension model as proposed by Brandt (2017) in Equation 15. Alternatively, linguistic competence can also be regarded as a multidimensional construct; hence, the item response data can also be fitted to the Rasch between-item multidimensionality MCML model as given in Equation 14 (Adams & Wu, 2007; Adams, Wilson, & Wang, 1997). The item response data can also be fitted to the default unidimensional Rasch model as presented in Equations 7 and 13 if linguistic competence is conceived as a single dimension.

Because the items in the current study are written from scratch, it is not possible to know how the items would behave until empirical data from the test administration are analysed; hence, Rasch analysis is needed to provide empirical evidence of the item properties and validity arguments of the construct. Fitting the item response data to the Rasch model in the administration stage thus concludes the current study.

2.6 Conceptual Framework

The theoretical framework in Figure 2.3 shows an interrelationship between the construct theory and the measurement theory. Specifically, the construct theory of linguistic competence specifies how the item response data can be fitted to the Rasch model; and in turn, the Rasch model provides the validity arguments for the construct of linguistic competence. Based on this interrelationship, the conceptual framework in Figure 2.4 is proposed. The top panel of the conceptual framework shows that, in the test, linguistic competence is operationalized as consisting of six subdimensions: graphology, lexical items, word classes, morphology, syntax, and semantics. Students' performance on the test may be influenced by their age, grade level, gender, ethnicity, native language, and geographical area. Therefore, it is important to ensure that the test items do not function differently across these demographic factors.

To test the conceptualization of linguistic competence in the top panel, there are three possible variants of the Rasch model where the item response data from the test can be fitted to. The bottom panel of the conceptual framework illustrates these three variants. The rightmost variant is the Rasch unidimensional model where there is only a single dimension underlying the item response data. The second variant is the Rasch between-item multidimensionality model where each item calls upon a single dimension, but the test is associated with six dimensions. The third variant is the Rasch

subdimension model, where linguistic competence can be decomposed into six subdimensions. Fitting the item response data to the Rasch model will provide empirical evidence of the conceptualization of linguistic competence.

2.7 Summary

Language test development is the synthesis of a language theory, a validation framework, and a measurement model. These are the three "almost magical procedures and formulae for creating the 'best' test" that are alluded to by Bachman and Palmer (1996, p. 3). However, as the reviews of past studies have shown, there are different procedures and formulae for developing and validating a language test, and that the mythical 'best' test never exists. Apart from the reviews of past studies, this chapter has also expounded on the different language theories, validation frameworks, and measurement models. The current study, however, did not subscribe to any one theory, framework, or model; but has taken an eclectic approach in portraying the theoretical and conceptual frameworks for the study. In the current study, the underpinning language theory of the diagnostic test is the construct theory of linguistic competence, one of the components of communicative competence. Linguistic competence is composed of six subdimensions, and each test item is written to tap into one subdimension only. The items are written to maximize the test usefulness, which is one of the most fundamental elements in all validation frameworks. To provide evidence of validity, the item response data is fitted to three variants of the Rasch model, namely the unidimensional model, the between-item multidimensionality model, and the subdimension model. The next chapter will illustrate in detail how the study would be carried out.

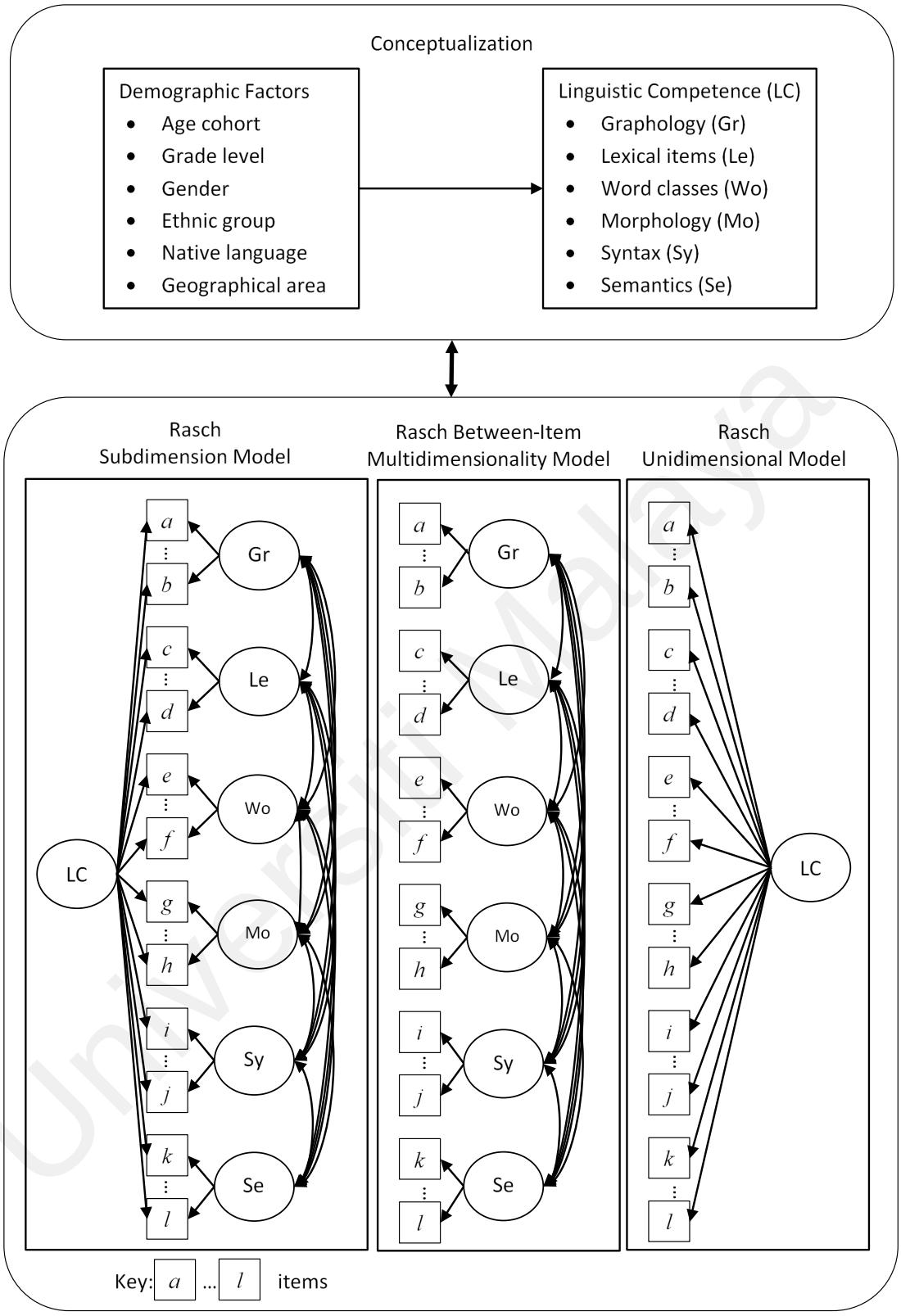


Figure 2.4. Conceptual framework of the study.

CHAPTER 3

METHODOLOGY

Theory is splendid but until put into practice, it is valueless (Penney, as cited in Mourdoukoutas, 2013).

The review of theories and past studies in Chapter 2 provides the theoretical basis upon which the current study is grounded. To put these theories into practice, Chapter 3 explains how the study is to be conducted so that the objectives outlined in Chapter 1 can be achieved. This chapter begins with a discussion of the research design and the procedural framework of the study. The intended population and the sampling for the study are also described, followed by the test design and operationalization. The chapter then proceeds with a plan on how the item response data can be analysed to address the research objectives put forth in Chapter 1 before ending with the results of the pilot study. The details of how the current study is conducted will allow for replicability in future research, which is of utmost importance since the current study concentrates on the early phase of the test development process.

3.1 Research Design

The sole focus of the study is on developing a diagnostic test of linguistic competence in the English language for lower secondary school students in Sarawak. From the theoretical framework in Figure 2.3, the final artefact of the test development process is item response data. Item response data are the test-takers' responses to the items in the test. This type of dataset is subjected to quantitative analyses as indicated in the discussion of measurement theories in Chapter 2. As such, all the research

questions asked are naturally framed as quantitative. In the current study, test administration occurs in the test-takers' natural setting, i.e. in their respective classrooms. This means that no manipulation of variables would take place; hence, the study employs a non-experimental quantitative research design (Johnson & Christensen, 2008). Specifically, a cross-sectional survey design is used, where the test is administered only at one point in time. A cross-sectional survey conducted on a representative sample is cost-effective, provides a rapid turnaround, and produces findings that can be generalized to the population (Creswell, 2012, 2014).

3.2 Procedural Framework

To elucidate the research design as discussed above, the procedure of the study is framed. Figure 3.1 illustrates the procedural framework for the study which shows the way the study will be conducted. The study begins with identification of the test purpose, conceptualization of the construct and writing of items to measure the construct. Since the purpose of the test is to provide diagnostic profile of lower secondary school student's strengths and weaknesses in the English language, the conceptualization of language knowledge must be related to specific areas which have been covered or will be covered in the immediate future. Therefore, to ensure that the diagnostic test can identify potential problematic areas of language knowledge, the relevant English language syllabi are drawn upon when writing the test items.

The initial pool of items is subjected to expert judgement to determine the degree to which the items are relevant to the construct being measured. At this stage, the expert judgement provides content-related validity evidence of the test. Based on the expert judgement, decisions are made to retain, revise, or reject the items. Items that have survived the expert judgment stage will form the diagnostic test. The test will

be administered to a representative sample from the population. Additional items on test-takers' demographic backgrounds will also be included. The data gathered at this stage will be analysed within the framework of Rasch model.

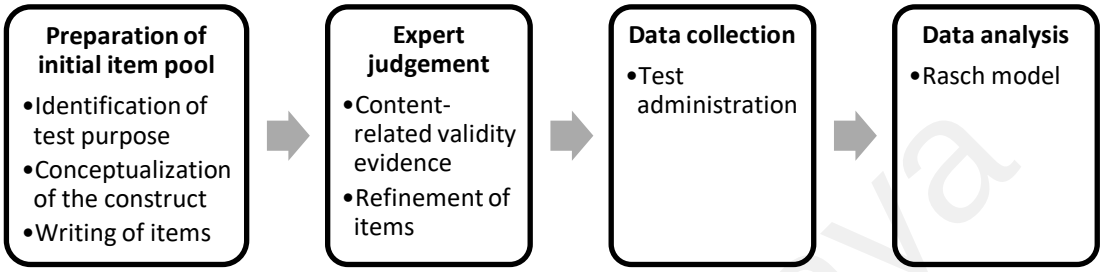


Figure 3.1. Procedural framework of the study. This framework is based on a combination of the models of test development that has been reviewed in Chapter 2.

3.3 Population and Sample

The target population is lower secondary school students in the state of Sarawak, Malaysia. This consists of students who are studying in Form 1 or Form 2 in national secondary schools that are under the purview of the Ministry of Education Malaysia. Their age range is from 13 to 15 years old. Form 3 students are excluded because of the restriction in the Ministry's consent for conducting research. Non-national secondary schools such as private schools and international schools are also excluded because it is not necessary for them to follow the national curricula. The target population is approximately 77,130 students spread across 182 schools (Sarawak State Education Department, 2017), within a geographical area of 124,450 km^2 that spans over 750km (Sarawak Convention Bureau, 2018). Drawing a representative sample from such a large and widespread population is beyond the scope of the current study, hence the study will be carried out in the southern zone of Sarawak. The state of Sarawak can be divided into three zones; and the southern zone

of Sarawak is the most diverse with a total population of approximately 1,172,618 (Sarawak Government, 2017). In other words, the southern zone can be perceived as a microcosm of Sarawak.

Based on the data obtained from the Sarawak State Education Department (2017), the southern zone of Sarawak encompasses 10 districts with 76 schools and a total of 17,860 Form 1 students and 17,298 Form 2 students. This constitutes an accessible population of 35,158 lower secondary school students ($N = 35{,}158$). For the sample to be representative of the population, the sample size must be sufficiently large (Fraenkel, Wallen, & Hyun, 2012). To determine the sample size required to estimate the main variable, which is the mean test score of the population, first the population variance must be estimated. Under the assumption that the test scores are normally distributed in the population with a range of 60 (because the test has 60 items), the population variance can be estimated using Deming's (1960) formula:

$$\text{Estimated population variance, } S^2 = 0.0289 \times (\text{range})^2 \qquad (16)$$

Therefore, $\text{Estimated population variance, } S^2 = 0.0289 \times 60^2 = 104.04$

Using $S^2 = 104.04$, the sample size required to estimate the mean test score with a bound, $B = 2$, on the error of estimation at the 95% confidence interval can be computed as follows (Scheaffer, Mendenhall, Ott, & Gerow, 2012):

$$\text{Sample size, } n_{\text{srs}} = \frac{NS^2}{(N-1)\left(\frac{B}{1.96}\right)^2 + S^2} \qquad (17)$$

Therefore, $\text{Sample size, } n_{\text{srs}} = \frac{35{,}158 \times 104.04}{(35{,}158-1)\left(\frac{2}{1.96}\right)^2 + 104.04} = 99.64 \approx 100$

At the 95% confidence interval, a simple random sample of 100 is required to estimate the mean test score of the population with a bound of ± 2 on the error of estimation, under the assumption that the test scores are normally distributed with a range of 60. As the population of lower secondary school students is widely dispersed

throughout the southern zone, it is impractical to use simple random sampling and spend an inordinate amount of time travelling to administer the test to 100 test-takers. Because schools form natural clusters, the cluster sampling method can be applied to select a representative sample from the accessible population (Kumar, 2011).

In educational settings, it is common for a well-planned probability sample design to accidentally turn into nonprobability sampling when the teaching staff of the sampled schools exercises subjective judgement in selecting students to participate in the study (Ross, 2005). To control for this bias, the current study will include all the elements in the sampled clusters. In other words, the study employs single-stage cluster sampling. It is pertinent to note that clustering increases sampling variation due to the homogeneity within the clusters. To reduce sampling variation, cluster sampling is often combined with stratification (Scheaffer et al., 2012). Stratification controls the distribution of a sample by ensuring representativeness in some important characteristics; and this can be achieved through explicit stratified sampling or implicit stratified sampling (Lynn, 2016). In a simulation study using real survey data, Lynn demonstrated that implicit stratified sampling provides better precision than explicit stratified sampling. Therefore, the test-takers for the study will be selected using a single-stage cluster sampling with implicit stratification.

The grade level of each school in the sampling frame forms an individual cluster. For example, all Form 1 students in School A is a distinct cluster from all Form 2 students in School A. Altogether, there are 152 clusters ($76 \text{ schools} \times 2 \text{ grades}$). Therefore, each cluster has an average size of $\bar{M} = \frac{35{,}158}{152} = 231$. The number of clusters required to achieve the same sampling accuracy as the simple random sample of 100 can be determined if the intra-class correlation is known. The intra-class correlation, which measures the degree of homogeneity within clusters, is used to

compute the design effect, which is the ratio of the variance of the estimate obtained from the complex sample to the variance of the estimate obtained from a simple random sample of the same size. For achievement outcomes such as test score, the default value of intra-class correlation is $\rho = 0.20$ (What Works Clearinghouse, 2008). Using the following formula from Ross (2005), the number of clusters can be determined as follows:

$$\text{Number of clusters}, n = \frac{n_{\text{srs}} \times [1 + \rho(\bar{M} - 1)]}{\bar{M}} \quad (18)$$

Therefore, $\text{Number of clusters}, n = \frac{100 \times [1 + 0.20(231 - 1)]}{231} = 20.35 \approx 21$

To achieve the same sampling accuracy as a simple random sample of 100, a total of 21 clusters must be sampled from the population of 152 clusters. To draw the 21 clusters, first, all the clusters are ranked according to grade level followed by district and urbanisation status. Within the urbanisation status, the ordering of the clusters is randomized. To improve the sample estimates of the population parameters, sampling with probabilities proportional to size is preferred over simple random sampling technique in cluster sampling (Scheaffer, Mendenhall, & Ott, 2006). Hence, systematic sampling with probabilities-proportional-to-size is applied to select the clusters. Each cluster is assigned a cumulative range according to the number of students in the cluster. For example, the first cluster with 290 students is assigned the range of 00001 to 00290; and the second cluster with 125 students is assigned the range of 00291 to 00415. Because $n = 21$ clusters are to be sampled, 21 numbers between 00001 and 35,158 must be selected. To select the 21 numbers, the sampling interval is calculated, $k = \frac{35,158}{21} = 1674.19$; and a random starting point between 0001 and 1674 is selected. The 21 numbers drawn are 893, 2567, 4241, 5916, 7590, 9264, 10938, 12612, 14287, 15961, 17635, 19309, 20983, 22657, 24332, 26006, 27680, 29354, 31028, 32703, and 34377. Clusters with cumulative range that contains any of these 21 numbers are

selected. This will result in an implicitly stratified sample of clusters. It is pertinent to note that each element in the cluster has equal probability of being selected. Finally, all students in the selected clusters are included in the sample. Figure 3.2 illustrates the distribution of the sample of clusters across the southern zone of Sarawak; while Table 3.1 summarises the number of clusters and elements in the sample and the population according to grade level, district, and urbanisation status. As can be seen from Table 3.1, the estimated sample size is 6,224 students or 17.70% out of 35,158 students in the population.

Figure 3.2. Distribution of the sample of clusters according to grade levels across the southern zone of Sarawak.

Table 3.1
Summary of Number of Clusters and Number of Students in Sample and Population Listed by Grade Level, District, and Urbanisation Status

Implicit Stratification Variables		Sample		Population	
		Number of Clusters	Number of Students	Number of Clusters	Number of Students
Grade Level	Form 1	11	3,730	76	17,860
	Form 2	10	2,494	76	17,298
	Total	21	6,224	152	35,158
District	Lundu	0	0	6	1,116
	Bau	2	666	6	2,081
	Kuching	6	1,426	48	10,526
	Samarahan	3	1,200	18	4,770
	Padawan	4	1,808	28	7,526
	Serian	2	467	12	3,111
	Simunjan	1	120	8	1,211
	Betong	1	193	14	2,319
	Lubok Antu	0	0	4	769
	Sri Aman	2	344	8	1,729
	Total	21	6,224	152	35,158
Urbanisation Status	Urban	7	2,085	56	13,099
	Rural	14	4,139	96	22,059
	Total	21	6,224	152	35,158

3.4 Instrument

The main instrument for the current study is a diagnostic test of linguistic competence in the English language for Form 1 and Form 2 students that aims to measure linguistic competence, which consists of six domains – graphology, lexical items, word classes, morphology, syntax, and semantics. Before the items are written, contents from the English language syllabi for Year 6 national primary schools, Year 6 national-type primary schools, and Form 1 (Ministry of Education Malaysia, 2015a, 2015b, 2017) were matched to the six domains. This is to ensure that the diagnostic test is based on specific areas of language knowledge which students have learned or will learn during English lessons (Alderson, 2005, 2007; Cumming, 2015). Following Fishman and Galguera's (2003) recommendation to begin with at least 15 items for

every domain of the test, the initial item pool for the test consists of 90 items (15 items $\times$ 6 domains).

Because a diagnostic language test should be discrete-point, user-friendly and efficient (Alderson et al., 2015), the most suitable type of test items appeared to be the multiple-choice format as it offers more flexibility, is simple to use and allows for precise interpretation which enhances the content-related validity of test scores (Osterlind, 2002). The suitability of the multiple-choice format to assess language knowledge especially in large-scale assessment is also supported by Fulcher and Davidson (2007), Kunnan (2008), and Morrow (1981). Therefore, all the 90 items will be written in the multiple-choice format. In writing the items, inspiration is drawn from English textbooks that are used in secondary schools in Sarawak, general grammar books, and linguistics reference books apart from the subject matter knowledge and teaching experience of the item writer. The items are written in such a way that they tap into knowledge, recognition, and/or manipulation of rules within the six domains. It is also important to note that each item only taps into one domain. Appendix D shows the specifications of the test.

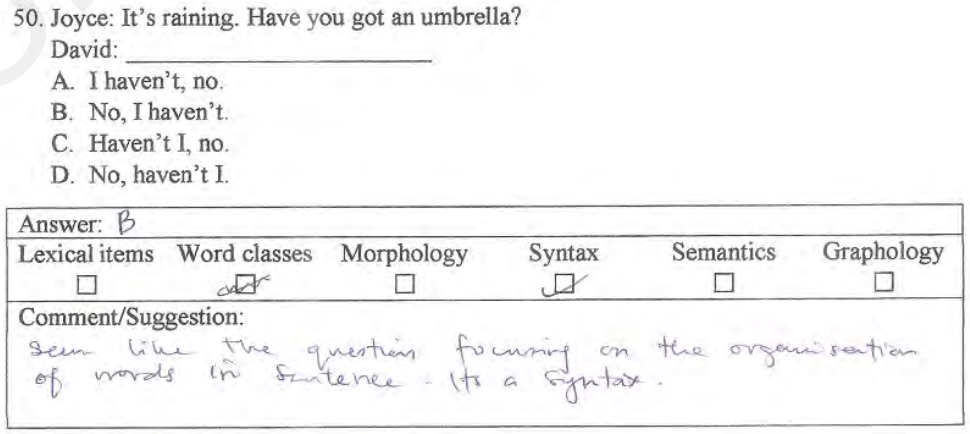
The items are then subjected to expert judgement. Specifically, seven experts from various backgrounds are recruited via personal contact. Table 3.2 shows the backgrounds of the experts. All the experts have been involved in English language education and educational testing in their career. Once the experts agree to participate in the study, they are given an informed consent form and the expert judgement form (see Appendix E). The expert judgement form consists of three parts: (a) an instruction sheet which defines the domains and explains how the judgement is to be done; (b) the list of test items (see Figure 3.3 for an example); and (c) a section on their personal details. As can be seen in Figure 3.3, the experts would have to read the item carefully,

tick the box corresponding to the domain that the item is testing, indicate the correct answer, edit the item if it is problematic, and give comments or suggestions to improve the item if necessary.

Table 3.2
Backgrounds of the Experts

Expert	Career	Academic Qualification	Current/ Pre-retirement Affiliation	English language teaching experience	Experience teaching lower secondary school students
1	Teacher (retiree)	▪ STPM	Secondary school	29	Yes
2	Teacher	▪ B. Ed. TESL	Secondary school	23	No
3	Teacher	▪ B. Ed. TESL	Secondary school	14	Yes
4	Teacher	▪ B. Ed. TESL ▪ M. Ed. TESL	Secondary school	6	Yes
5	Teacher	▪ B. Ed. TESL ▪ M. Ed. TESL ▪ PhD Education	Secondary school	25	Yes
6	Lecturer (retiree)	▪ BA ▪ M. Ed. (T & E) ▪ PhD (T & E)	Teacher training institute	15	No
7	Lecturer	▪ B. Ed. TESL ▪ MA Education ▪ PhD Education	Private university	6	No

Note. TESL = Teaching of English as a Second Language; T & E = Testing and Evaluation.


50. Joyce: It's raining. Have you got an umbrella?
David: _____
A. I haven't, no.
B. No, I haven't.
C. Haven't I, no.
D. No, haven't I.

Answer: B

Lexical items	Word classes	Morphology	Syntax	Semantics	Graphology
☐	☑	☐	☑	☐	☐

Comment/Suggestion:
Seem like the question focusing on the organisation of words in sentence - its a syntax.

Figure 3.3. Example of an expert's judgement of an item in the test.

The experts' answers to each test item are aggregated to determine if the item has problematic answer keys. Out of the 90 items, all the experts agreed with the answers to 71 items. For the remaining items, the percentage of agreement for the answers to 11 items was 86%, six items 71%, and the remaining two items at 57% and 29% respectively. This shows that the experts generally agreed with the answers to the items. For any item where a different answer is indicated by the experts, the answer options are revised. Similarly, the comments and suggestions given by the experts are also used to revise the items. Appendix F shows a summary of the expert judgement and the decision made to the items accordingly.

The domain specified by the experts for each item are also collated to check if the item could possibly tap into other domains. Only eight items had achieved 100% expert agreement. For the remaining items, the percentage of agreement was 86% for 36 items, 71% for 20 items, 57% for 15 items, 43% for nine items, and 29% for two items. The experts' opinions of the domain that the item is testing may provide some explanations to the results from the dimensionality analyses of the test. Depending on the dimensionality analyses on the item response data, some items may indeed tap into other domains than the one specified. At the early phase of test development, such items could be useful to model within-item dimensionality; hence, these items would not be removed. Instead, the disagreement between the experts and the item writer on the domains that the items are testing would be noted when conducting the dimensionality analyses.

Because Expert 2 and Expert 6 commented that a test with 90 items may cause fatigue for lower secondary school students, the set of test paper that each student receives only contains 60 items. Out of the 60 items, 54 items are common among all the sets while the remaining six items differ from set to set. Altogether, there are 6 sets

of the test. For each domain, the nine items judged to be least problematic are selected as the common linking items; and the remaining six items are assigned randomly to each of the sets. At the early phase of test development, it is important to investigate the properties of as many items as possible so that items that are problematic in terms of the response pattern can be discarded in later phases. The 54 common items are rearranged from seemingly easy to difficult within each domain; and the six additional items for each set are also arranged as such. It is pertinent to note that, at this point, the ordering of item difficulty is tentative and based on the judgement of the item writer. Each set of the test paper starts with items in the domain of graphology, lexical items, word classes, morphology, syntax, and semantics followed by the additional six items. To avoid confusing the test-takers, the test is not partitioned according to the domains. The instructions to the test is included in the cover page, which also explains the purpose of the study. The cover page and the 60 items constitute the test booklet. An answer sheet with items on demographic backgrounds is administered together with the test booklet. Both the cover page and the answer sheet are translated into the Malay language and are checked by a Malay language expert. The cover page and the answer sheet are therefore in bilingual.

3.5 Data Collection

Before any data can be collected, permission is first obtained from the Ministry of Education Malaysia through the Educational Planning and Research Division and the Sarawak State Education Department, which serves as the ethic committee overseeing any research conducted in national schools. When the letters of permission are obtained, letters of invitation to participate in the study are sent out to the principals of the schools in the sample. Attached with the invitation letter are the letters of

permission from the Ministry and an informed consent form. The informed consent form contains an information sheet with brief details of the study, certificate of consent if the school wishes to participate in the study, and a refusal-to-participate form if the school does not want to be involved in the study. To increase the participation rate, follow-up phone calls and/or courtesy visits to the schools are made. Suitable dates for the test administration are fixed and test administration procedures are discussed with the school principals. The school principals are assured that names of the schools would remain confidential and students would remain anonymous at all stages throughout the study.

On the date of the test administration, the teachers in the classroom who are invigilating the test are informed about the study and the test administration procedures. A pack containing the test booklets and answer sheets are handed over to the teachers. In their respective classrooms, the teachers distribute the test booklets and the answer sheets to the students. The students are expected to respond to the items in the test booklet by shading the corresponding answer options in the answer sheet. The students' responses to the test items constitute the main data to be collected in the study, i.e. the item response data. Information on the students' demographic backgrounds is also collected in the righthand panel of the answer sheet, where students are asked to check the appropriate boxes for their year of birth (to determine their age), gender, parents' ethnic groups, and native language. Information on the grade level and the geographical area is marked on the bundles of answer sheets when they are returned. During the test administration, the researcher walks around to brief the students about the study, thank them for their participation, and entertain any inquiries about the study. Students are given one hour and thirty minutes to respond to the test items and provide information on their demographic backgrounds. It is to be noted that the test is

administered under examination condition, where students do not discuss nor copy the answers. It is also important to note that different classes in the same school are randomly assigned different sets of the test booklet; and this information is marked on the bundles of answer sheets once they are collected from the teachers who invigilate the test.

As a token of appreciation and reimbursement for the time and effort in participating in the study, a letter of appreciation, a preliminary report of the study, and English books worth RM100 for the school library are sent to the schools involved within three months after the test administration has concluded. Appendix G displays the letters of permission from the Ministry of Education Malaysia, a sample letter of invitation to the school, the informed consent form, a sample of the test booklet, and the answer sheet.

3.6 Data Analysis

The item response data collected from the test-takers are fitted to the unidimensional Rasch model using Winsteps version 3.66.0 (Linacre, 2006); and to the Rasch multidimensionality and subdimension models using ConQuest version 4.14.2 (Adams, Wu, Macaskill, Haldane, & Sun, 2017). The following explains how the Rasch analyses are to be conducted to answer the research questions set out in Chapter 1.

Objective 1. To address the first objective, i.e. to assess the dimensionality of the test, the item response data are first fitted to the unidimensional Rasch model. The residuals that do not fit the Rasch model are subjected to principal component analysis (PCA), which looks for groups of items sharing similar patterns of unexpectedness. Items with the same patterns of unexpectedness most probably share a substantive

common attribute, and therefore, is an indication of a possible secondary dimension (Linacre, 2017). If the Rasch PCA of residuals reveals the possible existence of a secondary dimension, the item response data would be fitted to all the three variants of the Rasch model using a Monte Carlo approach to the calculation of integrals as implemented in ConQuest. The Monte Carlo approach is recommended for analysing data with three or more dimensions (Adams & Wu, 2010), as is the case for the current study where the data are expected to have six dimensions (see Figure 2.4).

Since the multidimensionality and the subdimension models are hierarchically related to the unidimensional model, the fit of the competing models can be compared using the differences in their deviance (G^2), where smaller deviance indicates greater likelihood of the solution, and thus closer fit of the estimated model to the true model (Baghaei, 2012). The difference in deviance between two competing models is approximately chi-square distributed with the difference between the number of parameters as degrees of freedom (Briggs & Wilson, 2003). Among the three competing models, the model with the smallest deviance and Akaike information criterion (AIC) is selected for further analyses.

Objective 2. To investigate the fit between the item response data and the Rasch model (Objective 2), the residual-based fit statistics for items are derived by squaring the standardised residuals and summing over persons. The standardised residual is the difference between the observed item response and the Rasch expected item response divided by the standard deviation of the item response. Wright and Masters (1982) proposed an unweighted (outfit) and a weighted (infit) fit statistic. The unweighted mean square is obtained by dividing the sum of the squared standardised residuals with the total number of respondents. As equal weight is given to all the standardised residuals, the unweighted mean square is relatively sensitive to

performance of persons that are not targeted by the specific items. To overcome this issue, the variance of the item response is used as weights in calculating the weighted mean square. This is because the variance of the item response is larger when an item difficulty is closed to a person ability than when they are distant.

When the item response data fit the Rasch model perfectly, the mean square is expected to have a value of one. The mean square can also be transformed to a standardised *t* statistic by taking the sample size into account. The *t* statistic can be interpreted as a normal distribution with an expected mean of zero and a standard deviation of one. This means that *t* statistic outside the range of ± 1.96 is an indication of misfit at the 95% confidence level. However, when the sample size is large enough, any minute misfit can be detected. Additionally, fit statistics are not absolute (Douglas, 1982). Therefore, Rasch analysts need to be careful about applying fixed limits of fit statistics when assessing the fit between the item response data to the Rasch model, which is essentially a "judgement call" (Wu et al., 2016, p. 148) and a "balancing act" (Bond & Fox, 2015, p. 86).

In the current study, the Bond-and-Fox developmental pathway in the form of bubble chart is plotted to visually assess the fit between the item response data and the Rasch model (Bond & Fox, 2015). The mean square range of 0.5 to 1.5, which is interpreted as productive for measurement (Linacre, 2017), is used as a guide. Any major misfits of items are investigated by removing the unexpected responses. The Rasch measures before and after removing the unexpected responses are cross-plotted to determine if the misfits are influencing the measurement. If the cross-plot is approximately a straight line, the unexpected response strings are not influencing the measurement; and hence the unexpected responses need not be removed. If there are measures that are noticeably off the diagonal identity line in the cross-plot, the misfits

are influencing the measurement; and hence the unexpected responses are removed during item calibration but reinstated for final reporting after anchoring the items at their calibrated locations (Linacre, 2015).

Objective 3. To estimate the reliability of the measures, both Winsteps and ConQuest reported test reliability based on classical test theory in the form of Cronbach's alpha or KR-20. However, they reported different reliability coefficients for the Rasch measures. For example, Winsteps reported the person and item separation and reliability indices (Linacre, 2017). Person separation is used to classify the test-takers into ability groups; while item separation is used to verify the hierarchy of item difficulty. The reliability index indicates whether the relative locations of the measures are reproducible. On the other hand, ConQuest reported the item separation reliability coefficient (without any separation index) and the expected a-posteriori (EAP)/plausible values (PV) reliability coefficient (Wu, Adams, Wilson, & Haldane, 2007). If the data is fitted to a multidimensionality model, ConQuest also reported the covariances and correlations between the dimensions. All reliability coefficients have a minimum of zero and a maximum of one, where high reliability suggests that the number of observations is sufficiently large, and the range is wide enough.

Objective 4. To determine how well the items are matched to the ability of the test-takers, the Wright item-person map is plotted. When the test is well-targeted to the ability level of the test-takers, the Wright map shows approximately bell-shaped curves that centre around the logit of zero for both the items and persons. There should not be large gaps between the items, and the range of item difficulty should approximately cover the range of person ability. Inspection of the Wright map can be used to inform the difficulty level of items that need to be added to the test, provided

that the sample of test-takers are representative of the test-takers that the test has targeted.

Objective 5. To examine whether the items function differently across the different age, grade, gender, ethnic, native language, and geographical groups, the differential item functioning (DIF) analyses are conducted. For an item to be classified as functioning differently across different subsamples, the difference in the item measures between the subsamples, i.e. the DIF contrast, must be larger than 0.50 logit to be noticeable, and the probability of observing this difference by chance must be less than .05 to be statistically significant (Linacre, 2017). For grade (Form 1 vs Form 2), gender (male vs female), and geographical (urban vs rural) groups, the Rasch item measures for each subsample are cross-plotted to determine if the items remain invariant within the modelled errors. Any item that is noticeably off the diagonal identity line in the cross-plot is indicative of failure of measurement invariance, a prima facie evidence of differential item functioning.

3.7 Pilot Study

A pilot study allows for preliminary analyses, which may give some indication of the tenability of the study and suggestion for further refinement (Ary, Jacobs, & Razavieh, 2002). The value of a pilot study cannot be overstated as Mackey and Gass (2012) pointed out, "no research project should be undertaken without extensive pilot testing" (p. 2). Therefore, to check the feasibility of the data collection and data analysis procedures, a pilot study was conducted on the Form 1 and Form 2 students of an urban school and a rural school in the district of Kuching. The two schools were selected at random from the remaining schools that were not in the sample. There were 159 students in the urban school and 441 students in the rural school that took part in

the pilot study ($n = 600$). Out of the 600 students, 313 of them were in Form 1 and 287 of them were in Form 2. Figure 3.4 shows the proportion of the sample according to demographic groups. The pilot study followed the data collection and data analysis procedures which were described earlier as closely as possible. The only difference in the data collection procedures between the pilot study and the main study, apart from the sample size, is the number of items in the test. Specifically, for the pilot study, the test consists of the 54 common items only. The data analysis procedures for the pilot study data were also simplified.

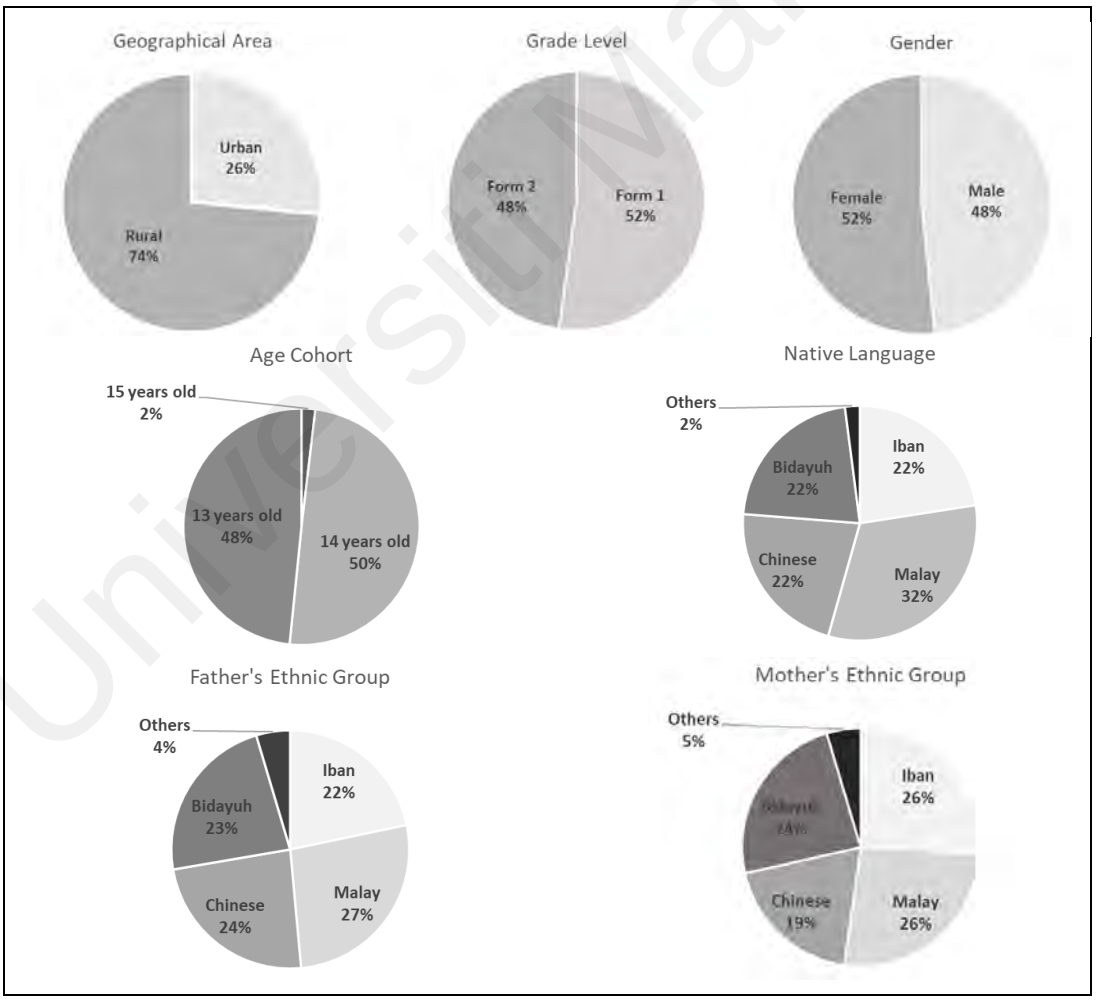


Figure 3.4. Distribution of pilot study sample according to geographical area, grade level, gender, age cohort, native language, and parents' ethnic group.

The item response data collected from the pilot study were fitted to the unidimensional Rasch model using Winsteps. Figure 3.5 shows the results of the Rasch PCA of residuals. The variance explained by the items (14.4%) was only three times the variance explained by the first contrast (4.8%), suggesting a noticeable secondary dimension in the items. Moreover, the eigenvalue of the first contrast was 3.6, indicating that the secondary dimension has a strength of more than three items. The contrast plot in Figure 3.5 shows that the three items at the top of the plot, i.e. Items 40 (A), 39 (B), and 42 (C), share the same content area. Specifically, these three items were designed to test the syntax domain. In contrast, the three items at the bottom of the loading, i.e. Items 50 (a), 49 (b), and 54 (c), belonged to the semantics domain. This means that the item response data collected using the test may not be unidimensional.

Since the Rasch PCA of residuals reveals possible multidimensionality, the item response data were also fitted to the between-item multidimensionality Rasch model and the subdimension Rasch model. Table 3.3 shows the global fit statistics for the three competing models. The multidimensionality Rasch model has the smallest deviance and the smallest AIC as compared to the unidimensional and the subdimension models. The change in deviance from the unidimensional to the multidimensionality model is statistically significant, $\chi^2(20) = 412.28, p < .001$. However, the change from the multidimensionality to the subdimension model is not significant, $\chi^2(12) = 0.25, p > .05$. This suggests that the item response data fit the multidimensionality model significantly better than the unidimensional model, but is comparable to the subdimension model. Because fitting the data to the subdimension model is computationally more complex than the multidimensionality model, the

remaining analyses would be conducted by fitting the item response data to the Rasch between-item multidimensionality model.

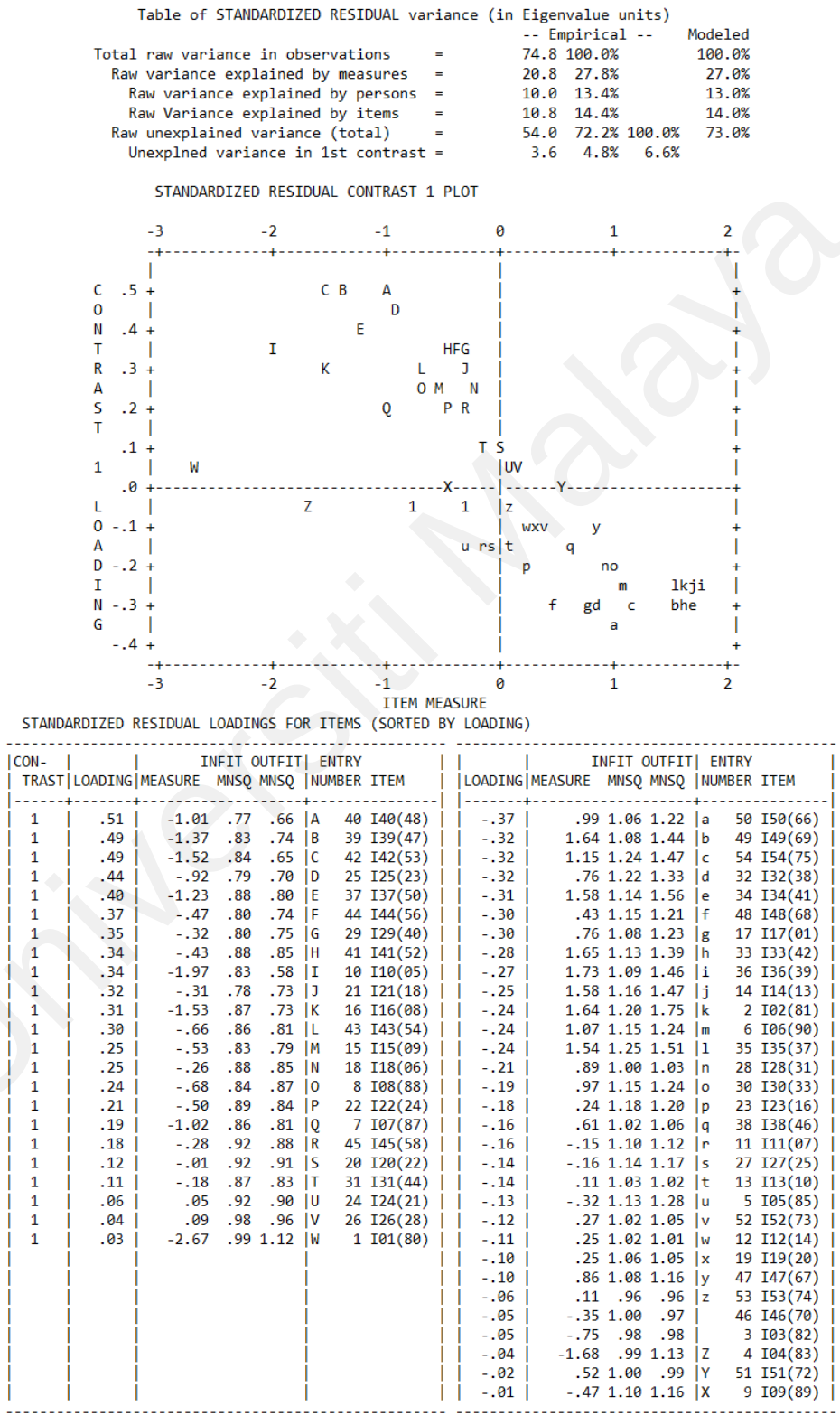


Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --		Modeled
Total raw variance in observations	=	74.8	100.0%	100.0%
Raw variance explained by measures	=	20.8	27.8%	27.0%
Raw variance explained by persons	=	10.0	13.4%	13.0%
Raw Variance explained by items	=	10.8	14.4%	14.0%
Raw unexplained variance (total)	=	54.0	72.2% 100.0%	73.0%
Unexplned variance in 1st contrast	=	3.6	4.8% 6.6%	

STANDARDIZED RESIDUAL CONTRAST 1 PLOT

STANDARDIZED RESIDUAL LOADINGS FOR ITEMS (SORTED BY LOADING)

CON-TRAST	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITEM
1	.51	-1.01	.77	.66	A 40	I40(48)	-.37	.99	1.06	1.22	a 50	I50(66)
1	.49	-1.37	.83	.74	B 39	I39(47)	-.32	1.64	1.08	1.44	b 49	I49(69)
1	.49	-1.52	.84	.65	C 42	I42(53)	-.32	1.15	1.24	1.47	c 54	I54(75)
1	.44	-.92	.79	.70	D 25	I25(23)	-.32	.76	1.22	1.33	d 32	I32(38)
1	.40	-1.23	.88	.80	E 37	I37(50)	-.31	1.58	1.14	1.56	e 34	I34(41)
1	.37	-.47	.80	.74	F 44	I44(56)	-.30	.43	1.15	1.21	f 48	I48(68)
1	.35	-.32	.80	.75	G 29	I29(40)	-.30	.76	1.08	1.23	g 17	I17(01)
1	.34	-.43	.88	.85	H 41	I41(52)	-.28	1.65	1.13	1.39	h 33	I33(42)
1	.34	-1.97	.83	.58	I 10	I10(05)	-.27	1.73	1.09	1.46	i 36	I36(39)
1	.32	-.31	.78	.73	J 21	I21(18)	-.25	1.58	1.16	1.47	j 14	I14(13)
1	.31	-1.53	.87	.73	K 16	I16(08)	-.24	1.64	1.20	1.75	k 2	I02(81)
1	.30	-.66	.86	.81	L 43	I43(54)	-.24	1.07	1.15	1.24	m 6	I06(90)
1	.25	-.53	.83	.79	M 15	I15(09)	-.24	1.54	1.25	1.51	l 35	I35(37)
1	.25	-.26	.88	.85	N 18	I18(06)	-.21	.89	1.00	1.03	n 28	I28(31)
1	.24	-.68	.84	.87	O 8	I08(88)	-.19	.97	1.15	1.24	o 30	I30(33)
1	.21	-.50	.89	.84	P 22	I22(24)	-.18	.24	1.18	1.20	p 23	I23(16)
1	.19	-1.02	.86	.81	Q 7	I07(87)	-.16	.61	1.02	1.06	q 38	I38(46)
1	.18	-.28	.92	.88	R 45	I45(58)	-.16	-.15	1.10	1.12	r 11	I11(07)
1	.12	-.01	.92	.91	S 20	I20(22)	-.14	-.16	1.14	1.17	s 27	I27(25)
1	.11	-.18	.87	.83	T 31	I31(44)	-.14	.11	1.03	1.02	t 13	I13(10)
1	.06	.05	.92	.90	U 24	I24(21)	-.13	-.32	1.13	1.28	u 5	I05(85)
1	.04	.09	.98	.96	V 26	I26(28)	-.12	.27	1.02	1.05	v 52	I52(73)
1	.03	-2.67	.99	1.12	W 1	I01(80)	-.11	.25	1.02	1.01	w 12	I12(14)
							-.10	.25	1.06	1.05	x 19	I19(20)
							-.10	.86	1.08	1.16	y 47	I47(67)
							-.06	.11	.96	.96	z 53	I53(74)
							-.05	-.35	1.00	.97	46	I46(70)
							-.05	-.75	.98	.98	3	I03(82)
							-.04	-1.68	.99	1.13	Z 4	I04(83)
							-.02	.52	1.00	.99	Y 51	I51(72)
							-.01	-.47	1.10	1.16	X 9	I09(89)

Figure 3.5. Rasch PCA of residuals (pilot study data).

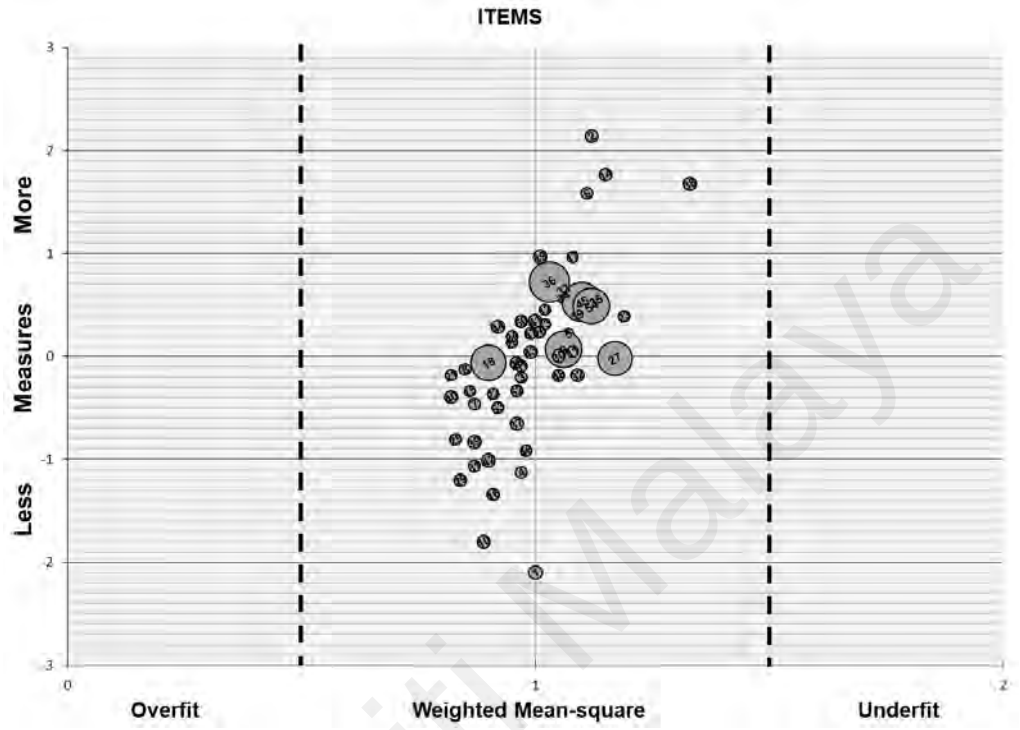
Table 3.3
Global Fit Statistics for the Three Competing Rasch Models (Pilot Study Data)

Model	Final Deviance, G^2	Change in G^2	Number of parameters	AIC
Unidimensional	36827.53	-	55	36937.53
Multidimensionality	36415.25	412.28	75	36565.25
Subdimension	36415.50	0.25	87	36589.50

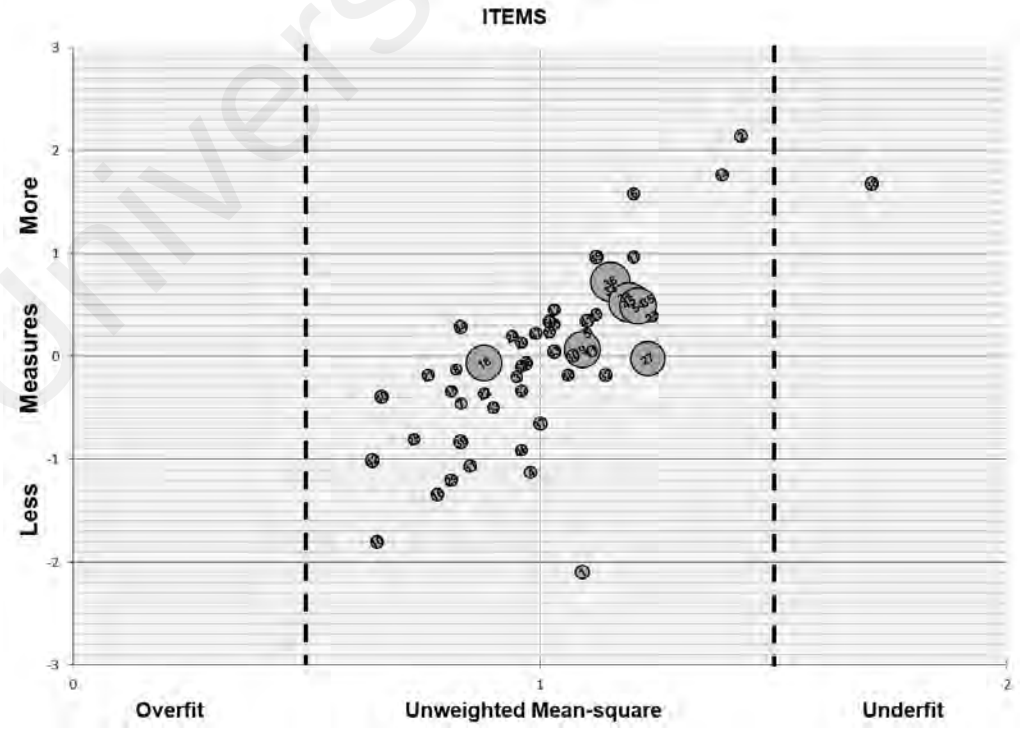
To investigate how well the item response data fit the multidimensionality model, the Rasch measures were plotted vertically on the logit scale and the mean-square fit statistics were plotted horizontally in the Bond-and-Fox developmental pathway, as illustrated in Figure 3.6. The area within the parallel dotted lines marked the pathway that is productive for measurement (i.e. between mean square range of 0.50 to 1.50). All the 54 items were located within the developmental pathway except for one item; Item 38 with a difficulty logit of 1.67 was located within the pathway in terms of weighted mean square (1.33) but outside the pathway in terms of the unweighted mean square (1.71). This means that, when the difficulty of Item 38 matched the test-takers' abilities, their responses fit the Rasch model, but for those whose abilities were not targeted by the item, their responses did not fit the Rasch model. This is verified by its item characteristic curve in Figure 3.7(a), which shows that test-takers with low ability had higher probability of success than modelled.

For the remaining 53 items, the developmental pathway suggests that the test-takers' responses did not deviate much from the Rasch model. For example, the item characteristic curve for Item 37, which has a weighted mean square of 0.96 and an unweighted mean square of 1.00, in Figure 3.7(b), shows that the empirical curve (dotted line) followed rather closely to the modelled curve (solid line). It can be concluded that the pilot study data fit the Rasch between-item multidimensionality model reasonably well. Any minor deviation from the model is not expected to cause

too much influence on the measurement. Therefore, unexpected responses such as those for Item 38 would not be removed at this stage.

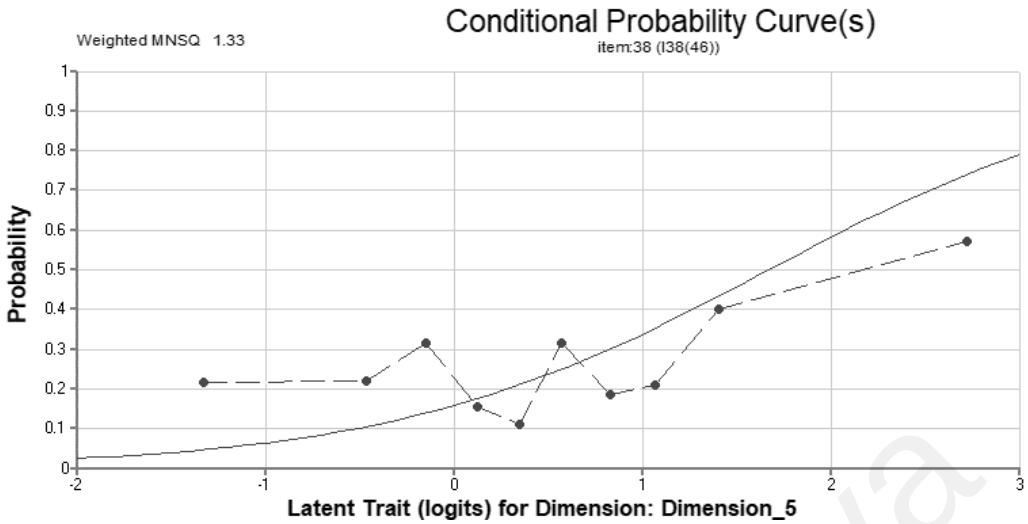


(a) Weighted mean square

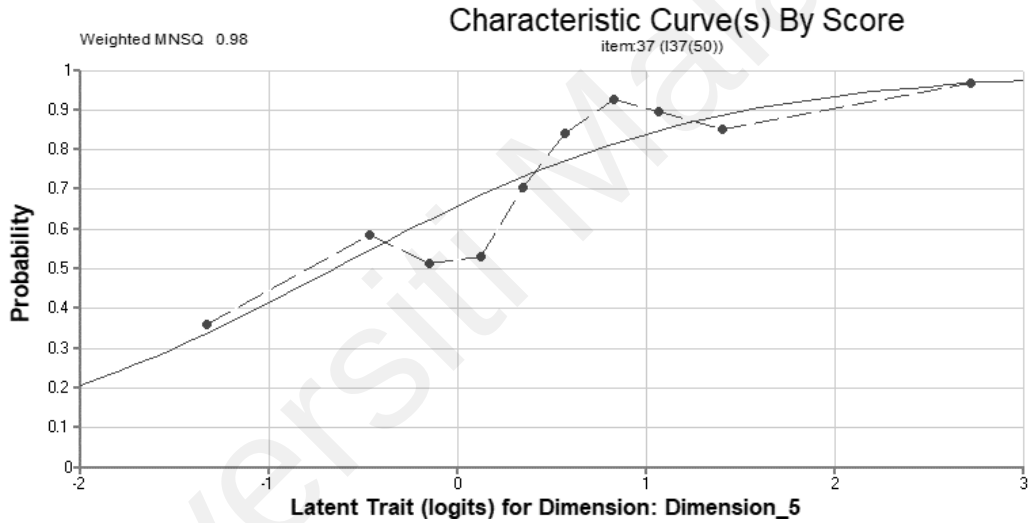


(b) Unweighted mean square

Figure 3.6. Bond-and-Fox developmental pathway for items (pilot study data).



(a) Item 38 with a logit of 1.67(SE = 0.07), weighted mean square of 1.33, and unweighted mean square of 1.71



(b) Item 37 with a logit of -0.65(SE = 0.07), weighted mean square of 0.96, and unweighted mean square of 1.00

Figure 3.7. Modelled and empirical item characteristic curves (pilot study data).

To estimate the reliability of the test and the Rasch measures, ConQuest reported that the KR-20 for the pilot study data was 0.88, suggesting that a large portion of the variance in the observed score is due to true score. The item separation reliability coefficient was 0.99, indicating that the sample of 600 test-takers was sufficiently large and diverse to verify the hierarchy of item difficulties. ConQuest also reported the reliability coefficient and variance for each dimension, as well as the correlations and covariances between the dimensions, which are summarized in Table

3.4. The reliability coefficients for the six dimensions ranged from 0.73 (semantics) to 0.87 (word classes), indicating that the Rasch measures for each of the dimension are reproducible. The estimated correlations between the dimensions ranged from 0.69 (between Graphology and Semantics) to 0.95 (between Lexical Items and Word Classes), implying that there are moderate to strong relationships between the six domains of linguistic competence.

Table 3.4
Covariances, Correlations, Variances, and Reliability Coefficients for Each Dimension (Pilot Study Data)

Dimension	Gr	Le	Wo	Mo	Sy	Se
Graphology (Gr)		0.734	0.805	0.416	1.163	0.438
Lexical Items (Le)	0.921		0.971	0.572	1.435	0.634
Word Classes (Wo)	0.919	0.948		0.631	1.640	0.664
Morphology (Mo)	0.747	0.878	0.882		0.894	0.450
Syntax (Sy)	0.845	0.892	0.928	0.795		0.934
Semantics (Se)	0.689	0.853	0.813	0.867	0.728	
Variance	0.681	0.933	1.125	0.455	2.778	0.592
EAP/PV Reliability Coefficient	0.787	0.863	0.871	0.752	0.860	0.730

Note. Values above the diagonal are covariances; and values below the diagonal are correlations.

To determine how well the test-takers were targeted by the items, the Wright map, as shown in Figure 3.8, is plotted. At first glance, it appears that the test-takers in the pilot study were reasonably well targeted by the difficulty of the items. First, the item distribution and the person distributions in most of the domains were spread out along the logit scale and peaked around the means with almost symmetrical tails, indicating that the distributions were approximately normally distributed. Moreover, the means of item measures and the person measures in three of the domains, i.e. Graphology, Lexical Items, and Word Classes, were located around the logit of zero. However, the means of the person measures for the domains of Morphology and Semantics were located around -1 logit, implying that the ability of the test-takers were

below those targeted by the items. Upon closer inspection, it was also found that the person distribution for the Syntax domain was platykurtic with some high ability test-takers not targeted by any item. The gap between the logits of 1 and 1.5 in the item distribution also implies that more items need to be added to the test.

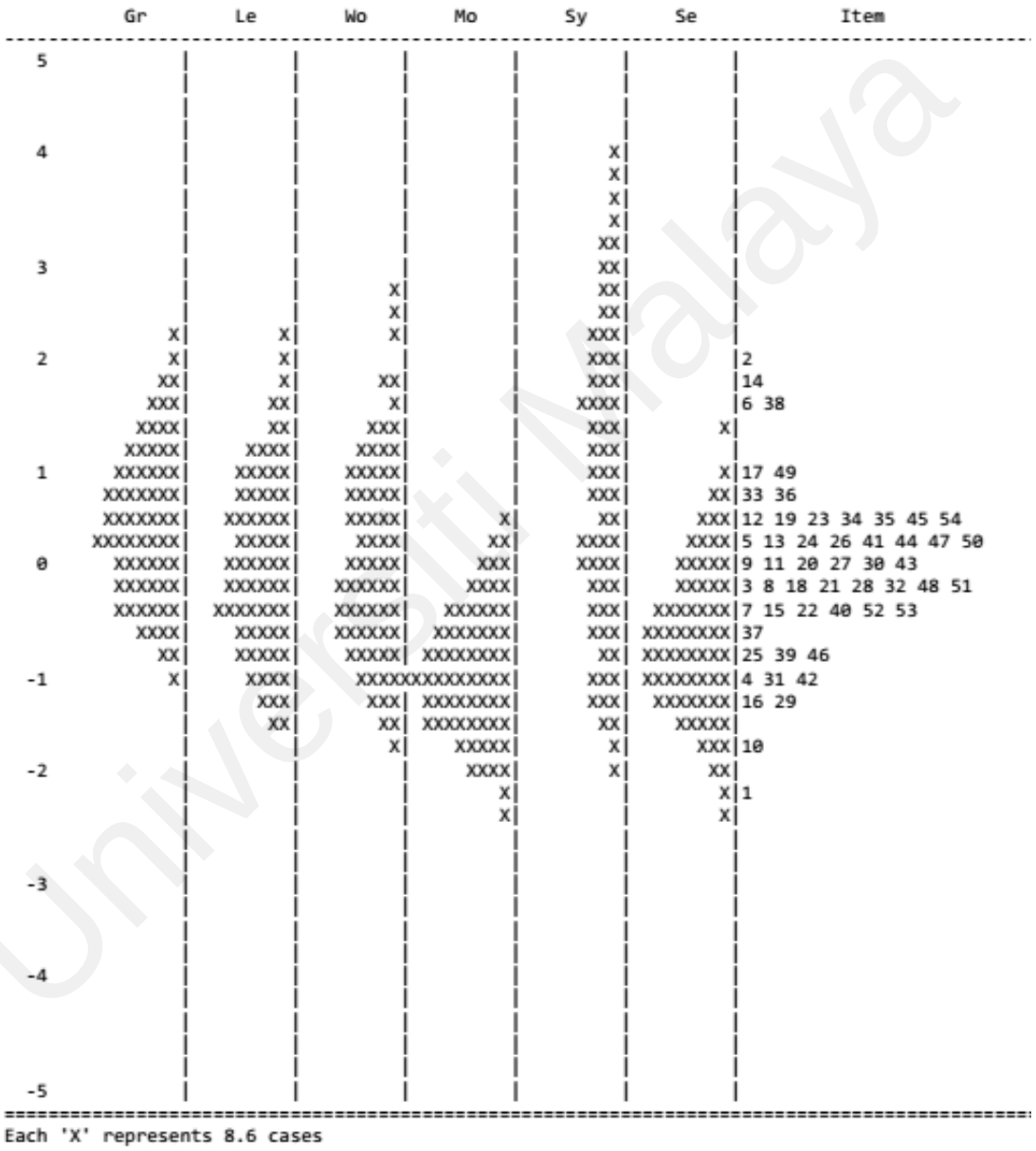


Figure 3.8. The Wright map (pilot study data).

3.8 Summary

The current study is conducted using a cross-sectional survey design on a representative sample of Form 1 and Form 2 students in the southern zone of Sarawak. As a balance between efficiency and practicality, the sample is selected using single-stage probabilities-proportional-to-size cluster sampling with implicit stratification. The sample of students responds to one of the six sets of 60-item diagnostic test of linguistic competence in the English language. In developing the test, 90 items in the multiple-choice format were written to tap into knowledge, recognition, and/or manipulation of rules within one of the six domains of linguistic competence, i.e. graphology, lexical items, word classes, morphology, syntax, and semantics. All the items have been revised according to recommendation from seven experts in the field of English language education and educational testing. The item response data collected using the test are fitted to the unidimensional, the between-item multidimensionality, and the subdimension Rasch models, before the further analyses were carried out using the best-fitting model. To demonstrate the feasibility of the study, a pilot study was conducted. Preliminary findings from the pilot study appeared to be promising.

CHAPTER 4

FINDINGS

Out of clutter find simplicity (Einstein, as cited in Wheeler, 1979).

To make some sense out of the data collected, the data are fitted to the Rasch model using the procedures outlined in Chapter 3. Chapter 4 reports the findings from the Rasch analyses, beginning with the dimensionality of the test. This is followed by an assessment of the fit between the data and the model, the reliability of the measures, and the match between the difficulty of the items and the ability of the test-takers. This chapter begins with the distribution of the test-takers according to the different demographic variables and ends with the DIF analyses across the different demographic groups.

4.1 Distribution of the Test-Takers

Chapter 3 proposed that 21 clusters with an estimated sample size of 6,224 students would be selected for the study. However, some of the schools were not able to participate in the study due to various constraints. The final sample of the study was composed of 16 clusters with 3,086 students. Figure 4.1 shows the distribution of the sample according to demographic groups. Out of the 3,086 students, 53% studied in Form 1 while the remaining 47% in Form 2. The ratio of Form 1 to Form 2 students in the sample was closed to the ratio in the population, which is approximately 50% (see Table 3.1). In terms of geographical area, 27% of the sample came from urban schools, which was 10% lower than the proportion of urban students in the population (see Table 3.1). There were slightly more female students (52%) than male students (48%)

in the sample. Half of the sample were 13-year-olds; 47% were 14-year-olds; and the remaining 3% were 15-year-olds.

The main ethnic groups in the sample were Malay, Iban, Bidayuh, and Chinese. Specifically, 26% of the students' fathers were Malays, 24% Iban, 23% Chinese, and 22% Bidayuh. In terms of mother's ethnicity, 26% were of Malay descendants, another quarter Iban, and a further 25% Bidayuh. Students with Chinese mothers only made up 19% of the sample. The different proportions between father's and mother's ethnicity indicate that some students came from interracial marriage. This suggests that the student's ethnicity may not reflect their native language. In fact, 31% of the students reported that they speak Malay as their mother tongue while 23% speak Iban. Bidayuh and Chinese speakers made up 20% of the sample respectively. The remaining 6% of the sample either speak other languages (such as English) or have more than one native language.

As outlined in the data collection procedure in Chapter 3, there were six sets of test booklet that were randomly assigned to the sample. Each set is composed of 54 link items that are common to all sets and six non-link items that are unique to each set. It is to be noted that the link items were placed in matching positions in the different sets, while the non-link items were placed at the end of the test. Out of the 3,086 students, 18% answered Set 4 and Set 5; 17% answered Set 1 and Set 3; while 15% answered Set 2 and Set 6 respectively. All the test-takers' responses for all the items were analysed concurrently in a single measurement framework.

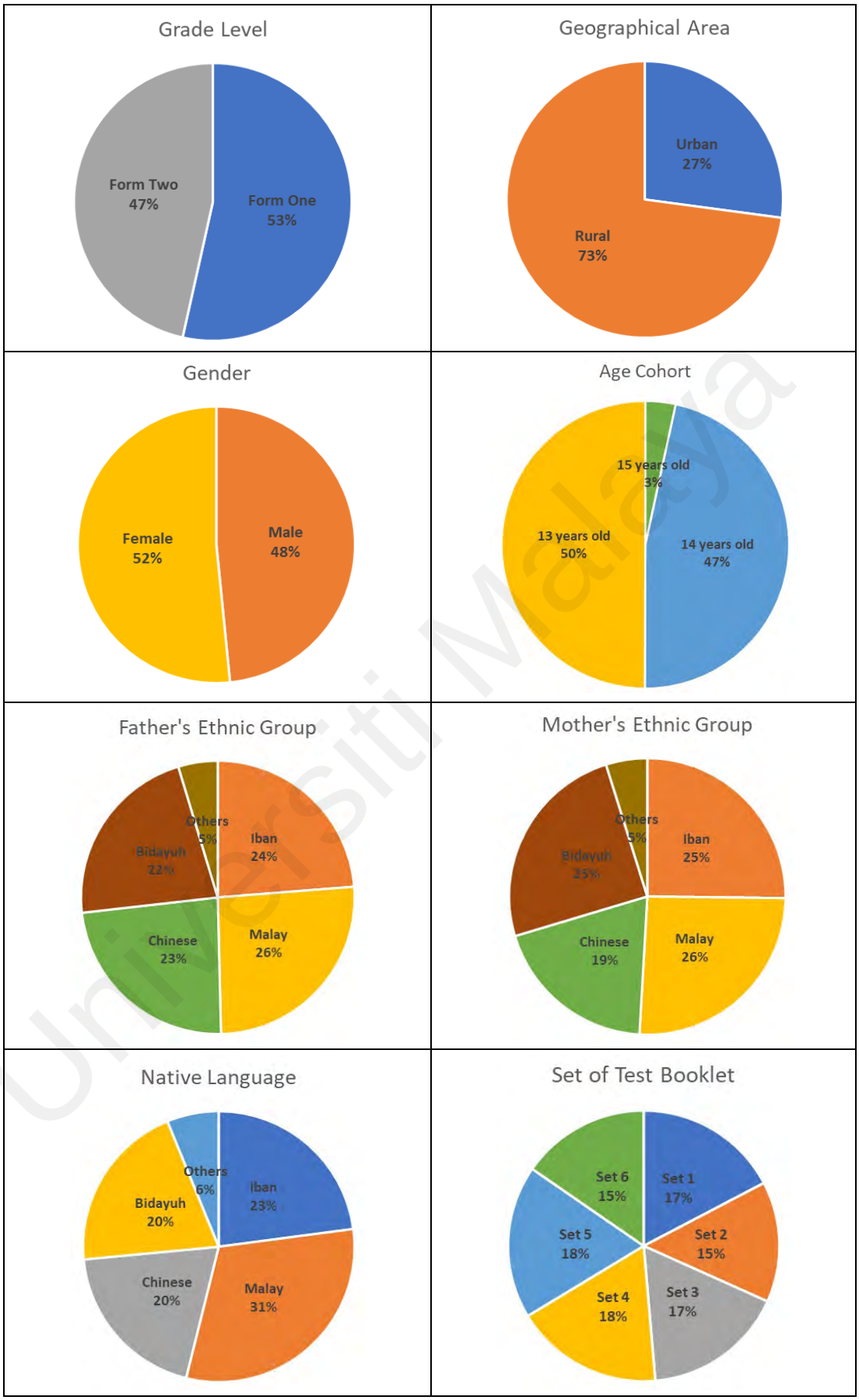


Figure 4.1. Distribution of the sample according to grade level, geographical area, gender, age cohort, parents' ethnic group, native language, and set of test booklet.

4.2 Dimensionality of the Test

The test-takers' responses to the 90 items were fitted to the unidimensional Rasch model using Winsteps. The residuals that did not fit the Rasch model were subjected to principal component analysis (PCA). Figure 4.2 shows the results of the Rasch PCA of residuals. The Rasch dimension explained 27.0% of the variance in the data, indicating that a large portion of the variance remained unexplained. The largest secondary dimension, as indicated by the first contrast in the residuals, explained 3.2% of the variance. Although the unexplained variance in the first contrast was only a quarter of the variance explained by the items (13.8%), the eigenvalue of the first contrast was 3.9. This is indicative of a noticeable secondary dimension and that the secondary dimension has a strength of about four items.

Figure 4.2 shows that the four items at the top of the contrast plot were Items 40, 39, 42, and 41; while the four bottommost items were Items 32, 35, 34, and 49. Appendix F reveals that the four topmost items were designed to test students' syntactical knowledge. In contrast, the three bottommost items, Items 32, 34, and 35, belonged to the morphology domain while Item 49 was intended to test the semantic domain. The difference in the shared contents of the items at the top and bottom of the plot suggests that the item response data collected using the test were not unidimensional. Since the Rasch PCA of residuals reveals possible multidimensionality, the item response data were fitted to the subdimension Rasch model and the six-dimensional Rasch model using ConQuest.

```
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Empirical --    Modeled
Total raw variance in observations     =       123.2 100.0%         100.0%
  Raw variance explained by measures   =        33.2  27.0%          26.1%
    Raw variance explained by persons  =        16.3  13.2%          12.8%
    Raw Variance explained by items    =        17.0  13.8%          13.3%
  Raw unexplained variance (total)     =        90.0  73.0% 100.0%   73.9%
    Unexplned variance in 1st contrast =         3.9   3.2%   4.4%

      STANDARDIZED RESIDUAL CONTRAST 1 PLOT

       -3          -2          -1           0           1           2
       -+-----------+-----------+-----------+-----------+-----------+- COUNT
        |                                   |                       |
C  .5 +              C B   A                |                       + 3
O       |                                   |                       |
N  .4 +                    F    DE          |                       + 3
T       |          I     H  G               |                       | 3
R  .3 +                      J     K        |                       + 2
A       |                N     OL   M       |                       | 4
S  .2 +                              QP     |                       + 2
T       |                      T  WUVRS     |                       | 6
   .1 +                    1   X Z1    Y1|                       + 6
1       | 1                      1111       |                       | 5
   .0 +--------------1----------------1-1-1121211-1----1--------------+ 14
L       |                      1   1       | 121 1 2 1           1 | 11
O -.1 +                            1       |       3 1             + 5
A       |                             u   t v   q    ysrw     xz   | 10
D -.2 +                                    |    m o np l           + 5
I       |                                   | j                k    | 2
N -.3 +                                    |        ie   hg f      + 5
G       |                                   |    a         bcd      | 4
  -.4 +                                    |                       +
       -+-----------+-----------+-----------+-----------+-----------+-
       -3          -2          -1           0           1           2
                                       ITEM MEASURE
 COUNT:  1         1  1  111 1112 116241642122315321322423321  32121   1

♠TABLE 23.3 Item Response Data.xlsx              ZOU701WS.TXT Jun  7 23:17 2018
INPUT: 3086 PERSONS  90 ITEMS  MEASURED: 3086 PERSONS  90 ITEMS  2 CATS   3.66.0
--------------------------------------------------------------------------------

 CONTRAST 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Empirical --    Modeled
Total raw variance in observations     =       123.2 100.0%         100.0%
  Raw variance explained by measures   =        33.2  27.0%          26.1%
    Raw variance explained by persons  =        16.3  13.2%          12.8%
    Raw Variance explained by items    =        17.0  13.8%          13.3%
  Raw unexplained variance (total)     =        90.0  73.0% 100.0%   73.9%
    Unexplned variance in 1st contrast =         3.9   3.2%   4.4%

  STANDARDIZED RESIDUAL LOADINGS FOR ITEMS (SORTED BY LOADING)
```

CON-TRAST	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ		ENTRY NUMBER	ITEM	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ		ENTRY NUMBER	ITEM
1	.52	-1.13	.78	.67	A	40	I40(48)	-.37	.54	1.23	1.34	a	32	I32(38)
1	.51	-1.49	.81	.70	B	39	I39(47)	-.36	1.29	1.29	1.60	b	35	I35(37)
1	.48	-1.61	.84	.67	C	42	I42(53)	-.36	1.42	1.20	1.60	c	34	I34(41)
1	.40	-.70	.83	.79	D	41	I41(52)	-.35	1.41	1.11	1.48	d	49	I49(69)
1	.39	-.63	.80	.74	E	44	I44(56)	-.32	.99	1.25	1.45	e	54	I54(75)
1	.38	-1.10	.81	.72	F	25	I25(23)	-.32	1.54	1.12	1.53	f	36	I36(39)
1	.36	-1.34	.85	.75	G	37	I37(50)	-.31	1.33	1.11	1.43	g	33	I33(42)
1	.36	-1.56	.84	.71	H	16	I16(08)	-.30	1.28	1.19	1.45	h	14	I14(13)
1	.33	-2.04	.86	.66	I	10	I10(05)	-.28	.91	1.03	1.18	i	50	I50(66)
1	.30	-.95	.83	.79	J	8	I08(88)	-.23	.17	1.21	1.25	j	23	I23(16)
1	.28	-.44	.81	.76	K	29	I29(40)	-.23	1.45	1.20	1.48	k	2	I02(81)
1	.27	-.79	.86	.79	L	43	I43(54)	-.22	.90	1.17	1.29	l	6	I06(90)
1	.26	-.39	.83	.79	M	21	I21(18)	-.22	.36	1.16	1.22	m	48	I48(68)
1	.25	-1.21	.85	.78	N	7	I07(87)	-.22	.71	1.02	1.10	n	28	I28(31)
1	.24	-.77	.87	.84	O	15	I15(09)	-.21	.56	1.04	1.13	o	17	I17(01)
1	.21	-.45	.89	.85	P	45	I45(58)	-.18	.75	1.12	1.21	p	30	I30(33)
1	.21	-.49	.89	.89	Q	22	I22(24)	-.17	.46	1.02	1.07	q	38	I38(46)
1	.16	-.35	.85	.81	R	31	I31(44)	-.16	.98	1.18	1.38	r	81	I5_57(35)
1	.15	-.27	.88	.85	S	20	I20(22)	-.16	.91	1.13	1.30	s	82	I5_58(02)
1	.14	-.75	.83	.78	T	73	I4_55(78)	-.16	-.03	1.07	1.07	t	13	I13(10)
1	.14	-.46	.81	.77	U	74	I4_56(11)	-.16	-.33	1.12	1.16	u	27	I27(25)
1	.14	-.37	.92	.90	V	18	I18(06)	-.15	.16	1.09	1.11	v	19	I19(20)
1	.13	-.58	.87	.83	W	83	I5_59(57)	-.15	1.10	1.22	1.57	w	77	I4_59(45)
1	.11	-.79	.87	.83	X	80	I5_56(19)	-.14	1.55	1.26	1.77	x	72	I3_60(63)
1	.11	-.13	.88	.86	Y	24	I24(21)	-.13	.83	1.17	1.29	y	64	I2_58(34)
1	.09	-.61	.89	.86	Z	65	I2_59(55)	-.13	1.60	1.14	1.44	z	66	I2_60(64)
1	.09	-1.05	.89	.84		55	I1_55(76)	-.12	.68	1.08	1.12		47	I47(67)
1	.08	-.06	.88	.88		56	I1_56(26)	-.12	.82	1.23	1.45		88	I6_58(27)
1	.08	-.60	.94	.94		79	I5_55(86)	-.10	.68	1.06	1.14		68	I3_56(12)
1	.07	-.60	.93	.90		71	I3_59(51)	-.10	-.40	1.12	1.22		5	I05(85)
1	.07	-.79	.95	1.16		67	I3_55(79)	-.08	.73	1.06	1.14		78	I4_60(62)
1	.05	-2.82	.98	.95		1	I01(80)	-.07	.21	1.00	1.00		12	I12(14)
1	.04	-.84	.96	.97		3	I03(82)	-.07	-.48	1.08	1.14		11	I11(07)
1	.03	-.68	.98	1.02		85	I6_55(77)	-.07	.79	1.11	1.21		62	I2_56(84)
1	.02	-.27	1.00	1.00		60	I1_60(71)	-.07	1.93	.93	1.23		86	I6_56(43)
1	.02	.19	.97	.97		51	I51(72)	-.06	.64	.99	1.10		58	I1_58(36)
1	.01	.48	.87	.85		69	I3_57(32)	-.06	.25	1.01	1.07		89	I6_59(60)
1	.01	.05	.92	.93		76	I4_58(49)	-.05	.33	.99	1.01		90	I6_60(61)
1	.01	.16	.97	1.00		59	I1_59(59)	-.04	.59	1.01	1.04		75	I4_57(29)
1	.00	-.48	1.01	1.00		46	I46(70)	-.04	.13	.99	1.01		52	I52(73)
								-.03	.48	1.01	1.05		61	I2_55(17)
								-.03	-.73	1.06	1.11		9	I09(89)
								-.02	-1.86	1.01	1.21		4	I04(83)
								-.02	.28	1.07	1.12		70	I3_58(30)
								-.02	-.17	1.12	1.29		87	I6_57(04)
								-.02	-.03	.92	.92		57	I1_57(03)
								.00	-.03	.98	.99		26	I26(28)
								.00	-.11	.93	.92		53	I53(74)
								.00	.85	1.04	1.07		63	I2_57(15)
								.00	.27	1.04	1.08		84	I5_60(65)

Figure 4.2. Rasch PCA of residuals.

Table 4.1 shows the global fit statistics for the unidimensional, subdimension, and multidimensionality Rasch models. Among the three competing models, the six-dimensional model has the smallest deviance and the smallest AIC, followed by the subdimension model. The change in deviance from the unidimensional model to the subdimension model is statistically significant, $\chi^2(24) = 2188.74, p < .001$. Similarly, the change in deviance from the unidimensional model to the six-dimensional model is statistically significant, $\chi^2(20) = 2290.07, p < .001$, and from the six-dimensional model to the subdimension model is also statistically significant, $\chi^2(4) = 101.33, p < .001$. The results indicate that the item response data fit the subdimension model significantly better than the unidimensional model, but the multidimensionality model provides the best fit. This suggests that the test measures six related unidimensional latent variables.

Table 4.1
Global Fit Statistics for the Three Competing Rasch Models

Model	Final Deviance, G^2	Number of parameters	AIC
Unidimensional	211391.11	91	211573.11
Subdimension	209202.37	115	209432.37
Multidimensionality	209101.04	111	209323.04

4.3 Assessment of Fit between the Data and the Rasch Model

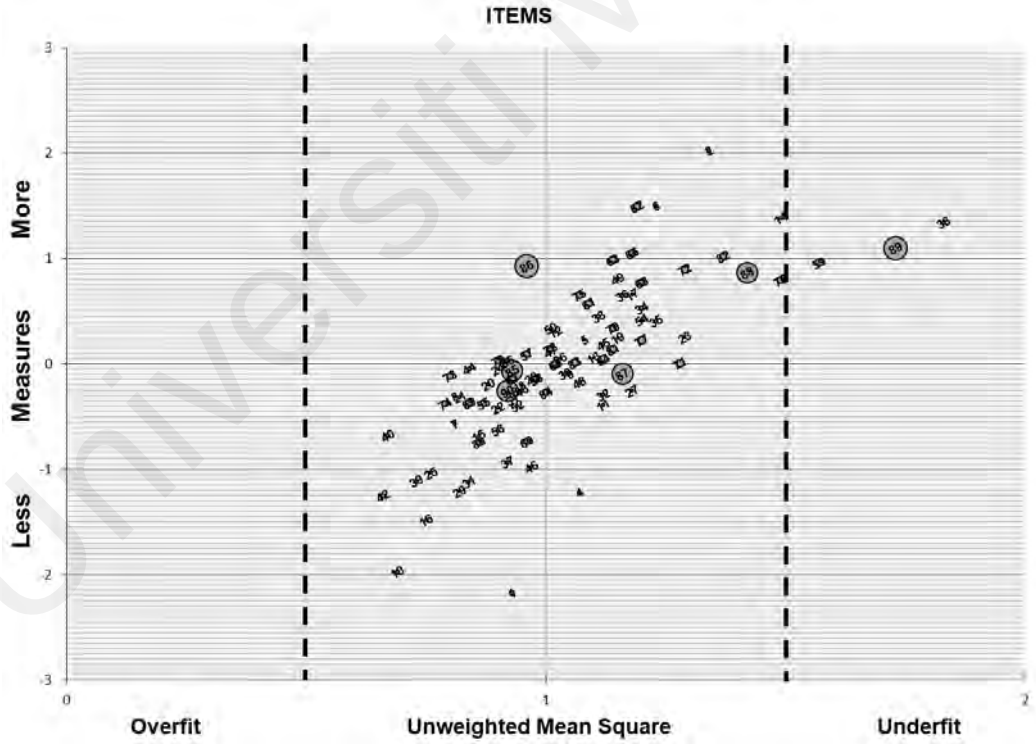
Since the item response data best fit the six-dimensional model, assessment of fit was conducted within the measurement framework of the between-item multidimensionality Rasch model using ConQuest. The Rasch measures were plotted vertically on the logit scale and the mean-square fit statistics were plotted horizontally in the Bond-and-Fox developmental pathway. The unweighted mean square for the items ranged from 0.66 to 1.83 with *t*-statistics ranging from -15.0 to 26.4; while the weighted mean square ranged from 0.83 to 1.37 with *t*-statistics ranging from -14.4 to

13.9. Figure 4.3 illustrates the Bond-and-Fox developmental pathway for items based on the item parameter estimates and fit statistics reported by ConQuest (see Appendix H). The area within the parallel dotted lines marks the pathway that is productive for measurement, i.e. between the mean square range of 0.50 to 1.50.

All the items were located within the developmental pathway except for three items: Items 38, 89 (59 of Set 6), and 59 (59 of Set 1). In terms of unweighted mean square, Item 38 with a mean square of 1.83 was located to the rightmost of the pathway, followed by Item 89 (1.73) and Item 59 (1.57). These three items, however, were located within the pathway in terms of weighted mean square. The fit statistics suggest that, when the difficulty of the items matched the test-takers' abilities, their responses fit the Rasch model; but when the test-takers' abilities were not targeted by the items, their responses did not fit the Rasch model.

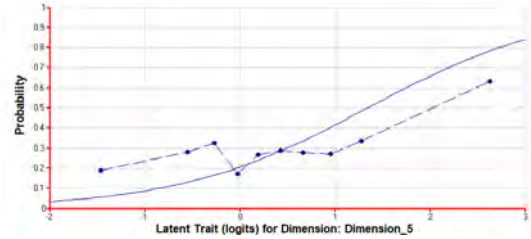
The above interpretation of fit statistics is verified by the item characteristic curves in Figure 4.4(a) to (c). The dotted line represents the empirical curve while the solid line represents the modelled curve. For the three underfitting items, the empirical curves started off above the modelled curves before they converged. For example, the dotted jagged line of Item 38 started off at a probability of .20 and continued to rise to .30 before taking a dip at logit zero. After this point, the dotted line began to converge with the solid line. Similar trend can be said of the other two items. The trend indicates that test-takers with low ability had higher probabilities of success than were modelled. Specifically, low achievers had approximately 20% chance of answering the items correctly, suggesting that guessing might be a factor causing the misfits. It is important to note that all the three underfitting items tested students' syntactical knowledge.

(a) Weighted mean square

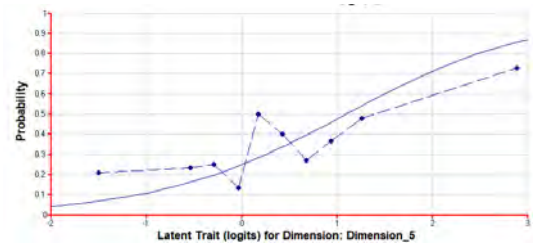


(b) Unweighted mean square

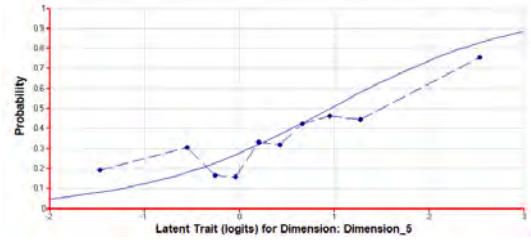
Figure 4.3. Bond-and-Fox developmental pathway for items.



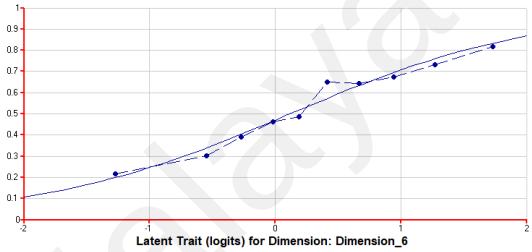
(a) Item 38 with a logit of 1.34 (SE=0.04), weighted mean square of 1.37, and unweighted mean square of 1.83



(b) Item 59 of Set 6 with a logit of 1.10 (SE=0.22), weighted mean square of 1.30, and unweighted mean square of 1.73



(c) Item 59 of Set 1 with a logit of 0.96 (SE=0.08), weighted mean square of 1.21, and unweighted mean square of 1.57



(d) Item 47 with a logit of 0.12 (SE=0.04), weighted mean square of 1.01, and unweighted mean square of 1.00

Figure 4.4. Modelled and empirical item characteristic curves.

For the remaining 87 items, the developmental pathway in Figure 4.3 suggests that the test-takers' responses did not deviate much from the Rasch model. Figure 4.4(d) shows the item characteristic curve for Item 47 with mean squares of 1. Unlike the empirical curves in Figure 4.4(a) to (c), the empirical curve for Item 47 followed rather closely to the modelled curve. This means that the probabilities of correct response for all levels of test-takers can be predicted by the Rasch model. Similar trend can be observed for the remaining 86 items. For these items, the item response data fit the between-item six-dimensional Rasch model.

For multidimensionality model, ConQuest also reports a person parameter estimate for each dimension and a single unweighted mean square as a person fit statistic that is computed using weighted likelihood estimates (Adams, 2010). The

unweighted mean squares for person ranged from 0.18 to 2.40. Figure 4.5 illustrates the Bond-and-Fox developmental pathway for persons based on the average case parameter estimates and unweighted mean square. Each circle represents a case and the size of the circle corresponds with the standard error of the parameter estimates. Figure 4.5 shows that the circles formed a lop-sided triangular shape with the majority located within the parallel dotted lines. This indicates that most of the cases fit the Rasch model. Some of the bigger circles were located to the top left of the pathway, suggesting that the responses from high performers overfit the Rasch model, but with a large margin of error. In contrast, a handful of smaller circles can be found towards the bottom right of the chart. This suggests that the responses from some low performers were more erratic than expected by the Rasch model. A plausible reason behind the erratic responses is guessing.

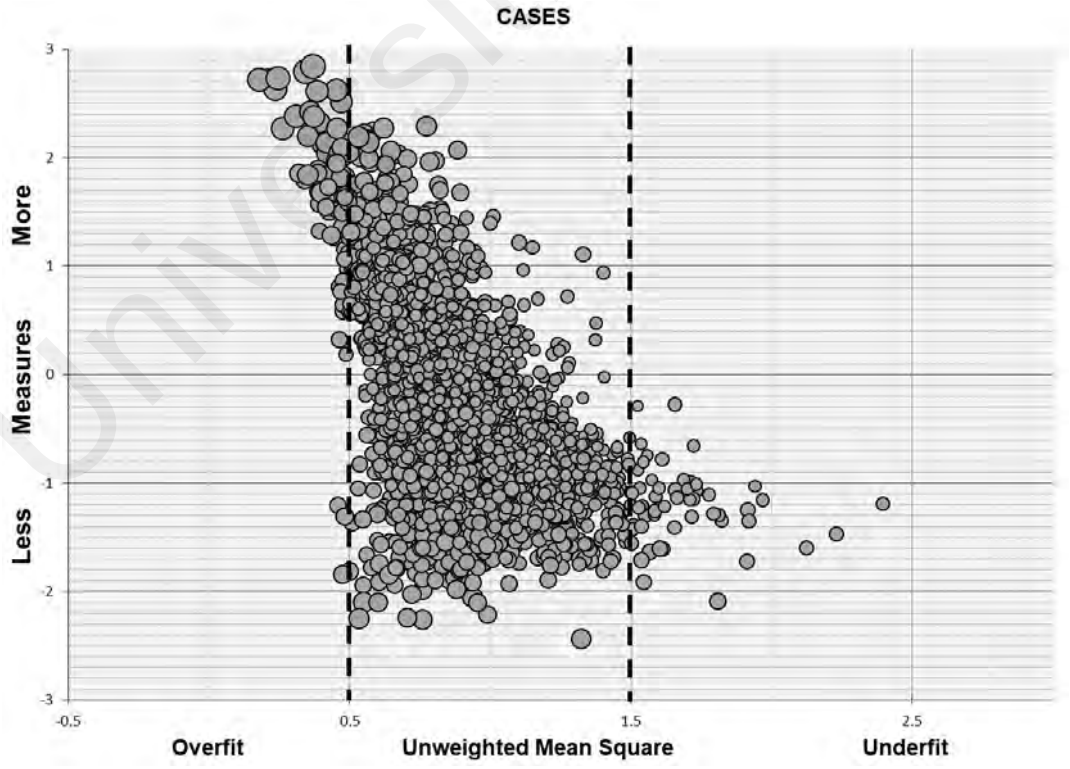
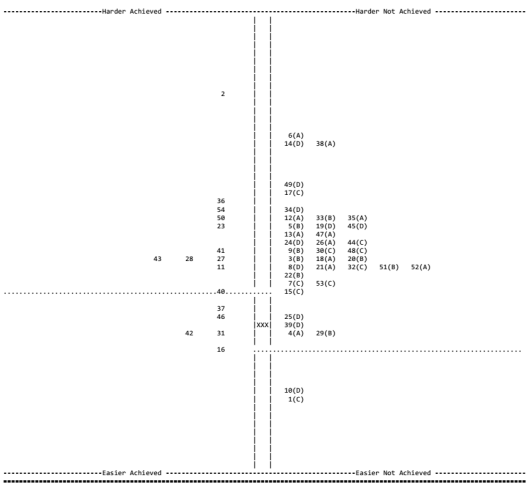


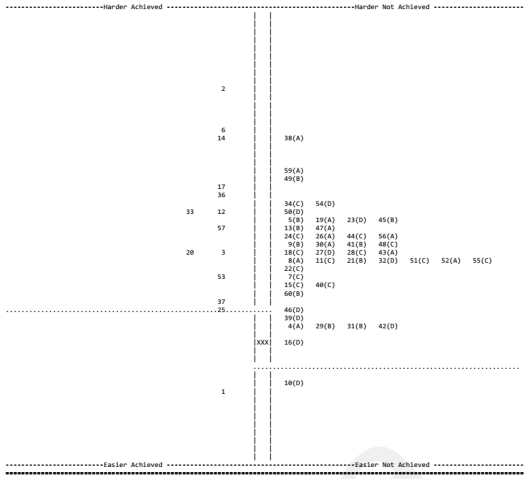
Figure 4.5. Bond-and-Fox developmental pathway for persons.

Guessing can be detected by careful examination of the kidmaps for underfitting cases. Figure 4.6 presents sample kidmaps from two underfitting persons, an overfitting person, and a case that fits the Rasch model to demonstrate the effect of guessing on the fit statistics. When the pattern of responses fits the Rasch model, it is expected that test-takers would be able to respond successfully to the easy items and incorrectly to the difficult items. For overfitting cases and cases that fit the model, most of the items should be in the bottom-left and top-right quadrants and very few items in the top-left and bottom-right quadrants. For example, Person 1127 with an ability of 1.84 and an unweighted mean square of 0.32 answered most of the easy items correctly (bottom-left quadrant) and none of the difficult items successfully (top-left quadrant). Meanwhile, Person 193 with an ability of -1.07 and an unweighted mean square of 1.00 answered most of the difficult items incorrectly (top-right quadrant) and none of the easy items incorrectly (bottom-right quadrant).

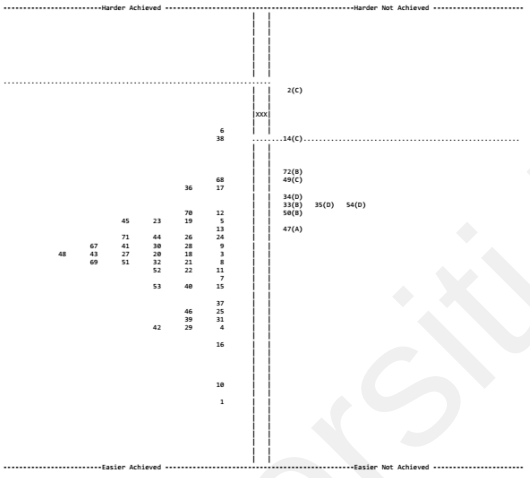
On the contrary, the pattern of responses for underfitting cases was more haphazard than the Rasch model would expect. Person 2314 with an ability of -1.19 and an unweighted mean square of 2.40 had unexpectedly responded to 11 difficult items correctly (top-left quadrant). Similarly, almost all the correct responses for Person 309 with an ability of -1.39 and an unweighted mean square of 1.51 belonged to the top-left quadrant. A careful examination of these difficult items reveals that the items came from various domains and were scattered along the logit scale. This suggests that Persons 2314 and 309 might have guessed the answers. Moreover, the percentage of correct responses for Person 2314 was 26.7% while that for Person 309 was 23.3%, which were expected if test-takers were to guess the answers at random for a test with four-option multiple-choice items.



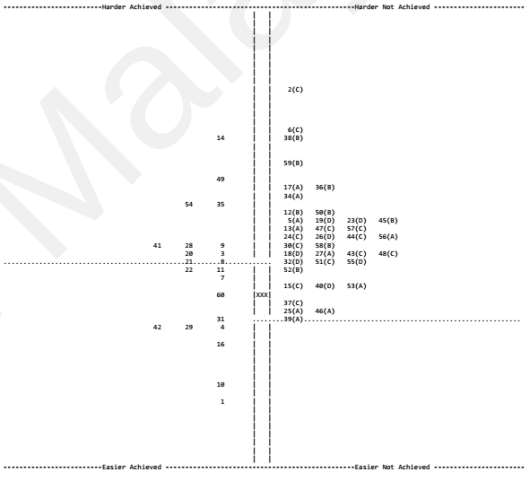
(a) Person 2314 with an average parameter of -1.19 (SE=0.77) and unweighted mean square of 2.40.



(b) Person 309 with an average parameter of -1.39 (SE=0.82) and unweighted mean square of 1.51.



(c) Person 1127 with an average parameter of 1.85 (SE=1.08) and unweighted mean square of 0.32



(d) Person 193 with an average parameter of -0.71 (SE=0.72) and unweighted mean square of 1.00

Figure 4.6. Sample kidmaps.

To minimize the effect of guessing on the item parameter estimates, all the underfitting cases were temporarily put aside for the item-calibration process. The underfitting cases represented only 5.83% of the sample. The item parameters before and after removing the underfitting cases were cross-plotted to determine if the misfits were influencing the measurement. Figure 4.7 shows the cross-plot of item parameters before and after removing the underfitting cases. The cross-plot was close to a straight line and none of the items was located outside the 95% confidence interval control

lines. This means that the presence of the underfitting cases were not influencing the item calibration. Therefore, the underfitting cases need not be removed during the item-calibration process.

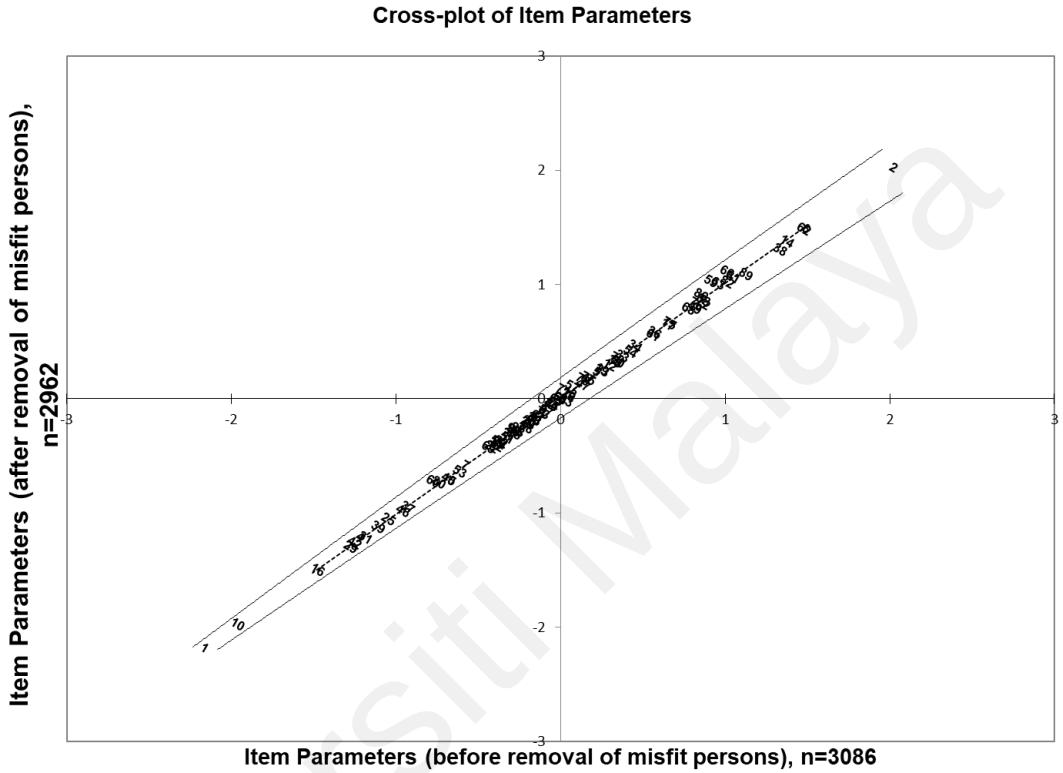


Figure 4.7. Cross-plot of item parameters before and after removing underfitting persons.

4.4 Item Discrimination

Apart from guessing, another possible factor behind the misfits was violation of the model assumption of item discrimination. Because item discrimination was held constant in Rasch model, ConQuest reported the point-biserial correlation under classical test theory as an index of item discrimination (Le, 2012). The relationship between point-biserial correlation discrimination estimates and Rasch fit statistics is almost monotonic except for the effect of item-person targeting on point-biserial ceilings (Wright, 1992). This nearly monotonic relation is demonstrated in the

scatterplot of the unweighted mean-squares against the point-biserial correlation in Figure 4.8. The grey circles represent items which had at least one distractor with positive point-biserial correlation, indicating the possibility of competing distractors. There were 27 such items, whereby 17 of them had item point-biserial below .20. Altogether, there were 21 items with point-biserial correlation of less than the recommended guideline of .20 (see Appendix I). Figure 4.8 shows that the item point-biserial discrimination ranged from -.05 to .57.

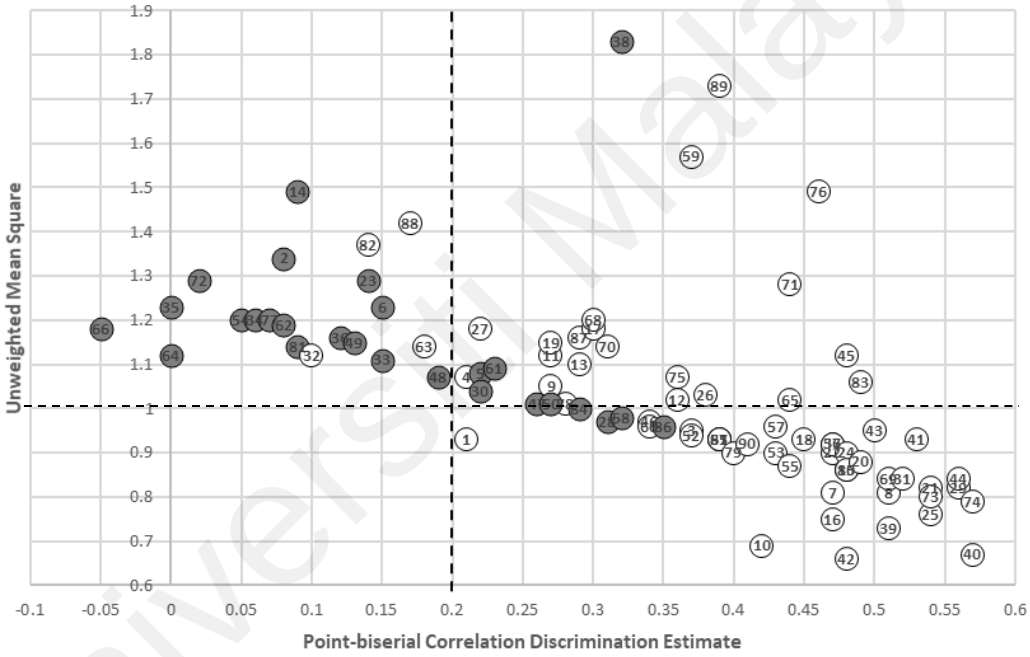


Figure 4.8. Scatterplot of unweighted mean-squares against point-biserial correlation discrimination estimates.

From the scatterplot in Figure 4.8, it can be determined that Item 66 (60 of Set 2) was the only item with a negative item point-biserial discrimination estimate (r_{pb}=-.05) with positive point-biserial for distractor D (r_{pb}=.21) and negative point-biserial for both distractor B (r_{pb}=-.02) and distractor C (r_{pb}=-.20). With a difficulty of 1.05 logit, this item was the most difficult item within the semantic dimension. Upon checking with the expert judgement (Appendix F), it was found that 14% of the experts

had selected option B and another 14% option D while the remaining 71% had selected the designated answer key A. Moreover, test-takers who had selected the answer key had mean ability of -0.83 (*SD*=0.54) for the semantic dimension, which was lower as compared to the mean ability of those who had selected distractor D (*M*=-0.74, *SD*=0.62). This means that test-takers with higher semantic ability tend to select distractor D over the answer key A. Option D is a competing distractor that could have qualified as an alternative answer. Because Item 66 was not wrongly keyed, it can be concluded that the observed responses to this item contradicted the general meaning of the test.

Among the items with possible competing distractors as represented by the grey circles in Figure 4.8, Item 86 (56 of Set 6) which tested students' morphological knowledge had the highest item point-biserial discrimination (r_{pb}=.35) but a positive point-biserial correlation for distractor C (r_{pb}=.13). However, test-takers who had selected the answer key D had higher dimensional mean ability (*M*=-0.21, *SD*=0.99) than those who had selected distractor C (*M*=-0.84, *SD*=0.74). This indicates that test-takers with the highest morphological ability tended to select the answer key D over the distractor C although those with relatively high morphological ability were attracted to distractor C. It is to be noted that Item 86 was the most difficult item within the morphology dimension. Therefore, Item 86 did not appear to be problematic; in fact, it was able to discriminate test-takers at the upper end of the ability continuum. Moreover, the item fitted the Rasch model relatively well with an unweighted mean square of 0.96 and a weighted mean square of 0.92.

From Figure 4.8, it is interesting to note that items with point-biserial discrimination of less than .20 had unweighted mean squares that were larger than the expected mean square of 1.00; and items with unweighted mean squares of 1.00 and

below had point-biserial discrimination above .20. There were 21 items in the former category and 41 items in the latter category. This indicates that items that are less discriminating do not fit the Rasch model very well, but items that fit the Rasch model as expected or better than expected are sufficiently discriminating. This also seems to suggest that most of the items in the test had point-biserial discrimination between .20 and .57; thus, items with discrimination outside this range would not fit the Rasch model well. In other words, the heterogeneity of item discrimination could be a factor behind the misfits. However, it is also pertinent to note that items with point-biserial discrimination above .20 had unweighted mean squares across the range. This means that items that can discriminate test-takers of different abilities relatively well may or may not fit the Rasch model, indicating that other factors such as guessing are at play in contributing to the misfits.

4.5 Reliability of the Item Placements and the Person Ordering

To estimate the reliability of the item placements and the person ordering along the logit scale, both Winsteps and ConQuest reported a variety of reliability indices. For example, Winsteps reported that the traditional coefficient alpha or KR-20 was .83, suggesting that a large portion of the variance in the observed score was due to true score. Winsteps also reported that the item separation had a lower bound of 12.01 and an upper bound of 12.37 with an item reliability of .99. Similarly, ConQuest also reported that the item separation reliability was .99. The high item separation and high item reliability imply that the sample of 3,086 test-takers in the study was sufficiently large and diverse to precisely place the items along the logit scale and establish a reproducible item difficulty hierarchy with 12 levels.

On the other hand, Winsteps reported that the person separation had a lower bound of 2.77 and an upper bound of 2.88 with a person reliability of .88. These indices suggest that the test was sensitive enough to precisely order test-takers along the logit scale and segregate them into two or three ability groups. These results were supported by the dimensional person separation reliability indices reported in ConQuest, as shown in the last row of Table 4.2. There were three dimensions with person separation reliability below .80: Morphology (.73), Semantics (.77), and Graphology (.78). This indicates that the range of items for these three domains might not be sensitive enough to distinguish between low and high achievers. For the remaining three dimensions, the highest person separation reliability index was .86 for the Word Classes domain, followed by Syntax (.85) and Lexical Items (.83). Since all the dimensions had the same number of items, the differences in the reliability indices suggest that certain dimensions might not have enough items across the range of test-takers' abilities. Nevertheless, the reasonably high item and person separation and reliability suggests that the relative locations of the items and persons are reproducible. Items and persons estimated to have high Rasch measures were more likely to have actual higher measures than those estimated with low measures.

Table 4.2 also shows the correlations and covariances between each dimension. The strongest correlation was between the Lexical Items and Word Classes dimensions (.94) while the weakest correlation was between the Graphology and Morphology dimensions (.70). There were no negative or weak correlation coefficients, suggesting that test-takers with high ability in one dimension most probably had high ability in the other dimensions. For example, a student with exceptional morphological knowledge is also expected to have good graphological knowledge. All the correlation coefficients were positive and strong, which were consistent with the directions

expected from a test of the same construct. Specifically, each of the six dimensions tested different aspects of linguistic competence; hence, it was not surprising that the dimensions were positively correlated to each other.

Table 4.2
Covariances, Correlations, Variances, and Reliability Coefficients for Each Dimension

Dimension	Gr	Le	Wo	Mo	Sy	Se
Graphology (Gr)		0.732	0.722	0.411	1.252	0.466
Lexical Items (Le)	0.897		0.884	0.556	1.489	0.609
Word Classes (Wo)	0.864	0.935		0.597	1.535	0.644
Morphology (Mo)	0.699	0.837	0.877		0.892	0.466
Syntax (Sy)	0.882	0.928	0.933	0.772		1.021
Semantics (Se)	0.747	0.863	0.890	0.916	0.832	
Variance	0.722	0.923	0.969	0.478	2.791	0.540
Person Separation Reliability	0.777	0.830	0.864	0.726	0.853	0.770

Note. Values above the diagonal are covariances; and values below the diagonal are correlations.

4.6 The Match Between Item Difficulty and Person Ability

To determine how well the test-takers were targeted by the items, the Wright map in Figure 4.9 was plotted. It mapped the distribution of the item parameter estimates against the spread of the person parameter estimates for each dimension. At first glance, it appears that the test-takers were reasonably well targeted by the difficulty of the items. First, the item distribution was spread out along the scale from a logit of -2.0 to +2.0 and peaked at the mean (zero logit) with two almost symmetrical tails. This indicates that the item distribution was approximately normally distributed. ConQuest reported that the item distribution has a skewness value of 0.38 and a kurtosis of -0.75, confirming the normal distribution of the item difficulty. Similarly, the person parameter estimates for each dimension were also spread out along the logit scale.

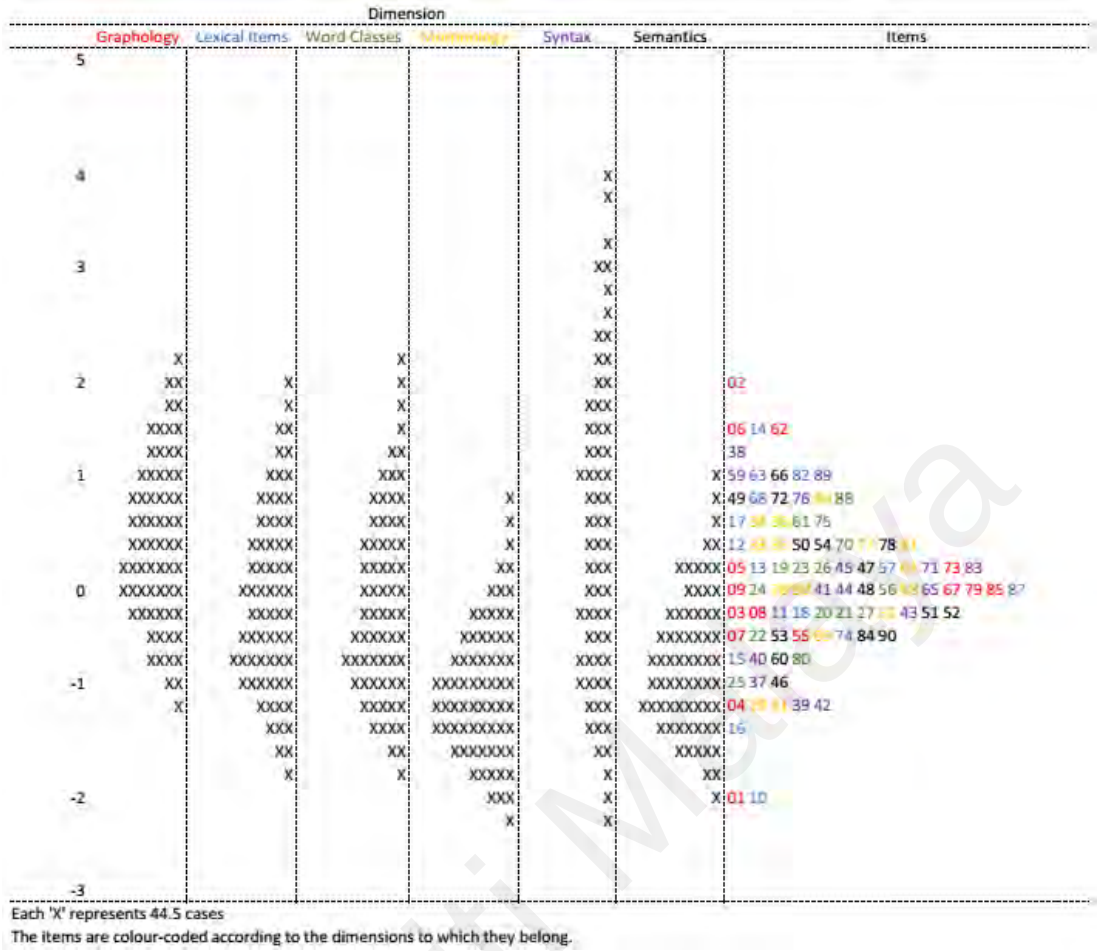


Figure 4.9. The Wright map.

However, upon closer inspection, it was found that only the test-takers' ability in the Graphology domain peaked at the mean of zero logit. For Lexical Items and Word Classes, the person parameter estimates peaked at above the logit of -1.0; while for the Morphology and Semantic dimensions, the estimates peaked at below the logit of -1.0. For these five dimensions, the tail in the direction of higher measures was slightly longer than the other tail, indicating that the distributions of test-takers' parameter estimates were slightly positively skewed. In other words, there were slightly more test-takers with lower ability estimates in terms of graphology, lexical items, word classes, morphology, and semantic knowledge. For the Syntax dimension, the distribution of person parameter estimates was platykurtic, spreading out from a

logit of -2.0 to +4.0. As compared to the other five dimensions, the person parameter estimates for the Syntax domain had the largest variance, i.e. 2.79 (see Table 4.2).

When the distribution of person parameter estimates for each dimension was matched to the items in the respective dimension, it was found that there were some gaps in the item distribution. For the Graphology dimension (red coding), the items covered the range of test-takers, but with gaps in between the logit of -1.2 and -0.6, and between 0.2 and 1.5. At the logit of -2.2, Item 1 was too easy for the test-takers. In contrast, for the Morphology (orange coding) and Semantics (black coding) dimensions, the items only covered the upper end of the scale from -1.0 to +1.0; hence, test-takers with ability lower than -1.0 were not measured by any of the items in these two dimensions. Furthermore, there was a gap in between the logit of -1.1 and -0.3 for the Morphology dimension. As for the Word Classes (green coding) and Syntax (purple coding) dimensions, the items measured test-takers in the middle of the range from about -1.0 to +1.0 with no items targeting test-takers of higher or lower abilities. The only dimension that covered the range of test-takers with little gaps in between was the Lexical Items domain (blue coding).

4.7 Differential Item Functioning Across Grade Levels

Overall, Form 1 students have performed more poorly than Form 2 students with a parameter estimate of -0.095 (SE=0.005). The actual parameter estimate for the Form 1 students was 19 times larger than its standard error estimate; thus, the mean difference of .19 between the Form 1 and Form 2 students was obviously significant. The chi-square value of 303.56 on one degree of freedom is consistent with this finding. Although the Form 1 students' mean performance was significantly lower than that of the Form 2 students, it did not indicate differential item functioning (DIF).

To provide prima facie evidence of DIF, the item parameters estimated from the Form 2 subsample were plotted against those from the Form 1 subsample. Figure 4.10 shows the cross-plot of item parameter estimates across grade levels. From the cross-plot, it is apparent that the locations of the measures for Items 1, 55, 69, 74, and 77 varied across grades by more than the modelled errors; while Items 10, 29, 65, 67, 72, 73, and 89 were located at the border of the error band. These items were expected to exhibit DIF. The rest of the items appeared to remain invariant within the modelled errors. The interactions between the item and grade facets were found to be statistically significant, $\chi^2(84) = 304.63, p < .001$. This confirms the existence of DIF in the items.

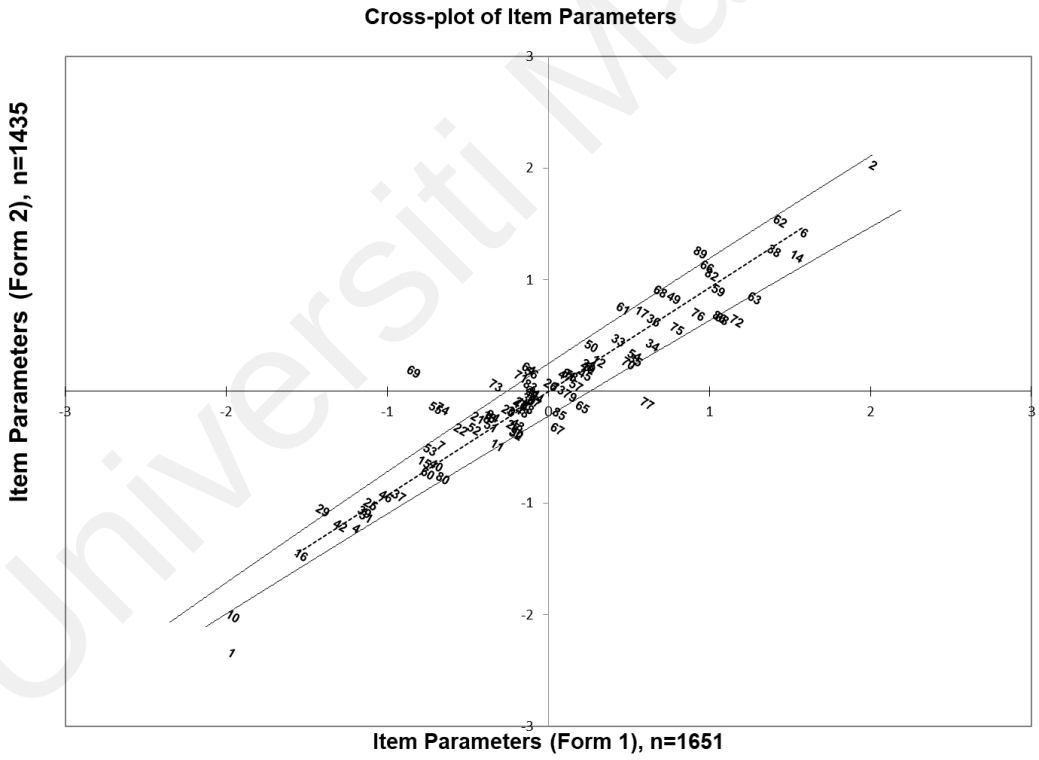


Figure 4.10. Cross-plot of item parameters for Form 2 subsample against Form 1 subsample.

Appendix J(i) shows a summary report of the DIF analysis. The report shows that 22 items were relatively easier for Form 1 than Form 2 students. For example, Form 1 students found Item 1 to be easier by a logit of 0.40 as compared to Form 2

students; thus, the estimate of 0.20 must be subtracted from the difficulty of this item for Form 1 students and 0.20 must be added for Form 2 students. On the other hand, Form 2 students found 28 items to be relatively easier than did Form 1 students. This indicates that at most 50 items would exhibit DIF. It is to be noted that the items that were noticeably off-diagonal in the cross-plot in Figure 4.10 were included in the list of 50 items that were most likely to exhibit DIF. As for the remaining 40 items, both Form 1 and Form 2 students found them to be equally difficult. The DIF contrast for all the items ranged from 0.004 to 1.022.

While the above analysis has shown the existence of DIF in the 50 items, it is the magnitude of the DIF that will determine if its effect is of substantive importance. For example, Item 25 is more difficult for Form 1 than Form 2 students, but the difference is only 0.10 logits. To objectively determine if the DIF is of substantive importance, the Mantel-Haenszel method is used. The magnitude of the Mantel-Haenszel statistics ranged from 0.001 to 1.581. Except for the item with the largest Mantel-Haenszel statistics, the Mantel-Haenszel statistics for the remaining 89 items were below 1.00 with non-significant chi-squares on four degrees of freedom. This indicates that the 89 items had negligible DIF.

In contrast, Item 69 with a Mantel-Haenszel statistic of 1.581 (p=.014) exhibited a moderate to large DIF. Specifically, Form 2 students found Item 69 to be easier than Form 1 students by 1.02 logits. This is reflected in the item characteristic curve in Figure 4.11 below. The empirical curve for Form 2 students was consistently above that for Form 1 before taking a dip at 0.50 logits. This shows that Form 2 students with ability below 0.50 logits found Item 69 to be relatively easier than Form 1 students with comparable ability. When both empirical curves levelled off, students from both grades found the item to be equally easy. Item 69, i.e. Item 57 of Set 3,

tested students' knowledge on the morpheme '-ess' for the word 'stewardess'. A plausible reason behind the DIF is that the use of this morpheme was not covered in the Form 1 syllabus but has been taught in the Form 2 syllabus; therefore, Form 1 students of lower ability have lower probability of getting this item correct than Form 2 students of lower ability. Form 1 students of higher ability might have prior knowledge of this morpheme and thus have equal chance of answering this item correctly as Form 2 students of higher ability.

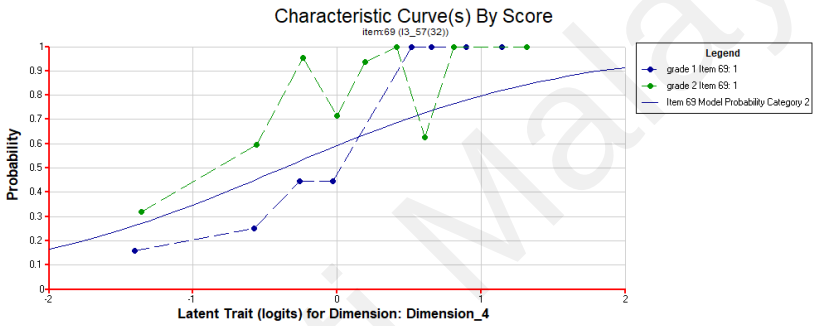


Figure 4.11. Empirical curves for Item 69 by grade levels.

4.8 Differential Item Functioning Across Age Cohorts

Although the items were tested on two grade levels, there were three age cohorts in the sample. Some students could be one year older than their peers in the same grade level due to various reasons such as an additional year of schooling before entering Form 1. The three age cohorts were the fifteen-year-olds, fourteen-year-olds, and thirteen-year-olds. Overall, it was found that the thirteen-year-olds performed better than the fourteen-year-olds, but the fifteen-year-olds recorded the best performance. The actual parameter estimates for the thirteen-, fourteen-, and fifteen-year-olds were -0.103 (SE=0.009), -0.276 (SE=0.006), and 0.379 (SE=0.008) respectively. The differences in the performance between the age cohorts were statistically significant, $\chi^2(2) = 4943.95, p < .001$.

Similarly, the interactions between the item and age facets were statistically significant, $\chi^2(168) = 4628.85, p < .001$, indicating the existence of DIF in the items. Appendix J(ii) shows a summary report of the DIF analysis for age cohorts. The report shows that 35 items were relatively easier for the fifteen-year-olds than the other two age cohorts. For example, the estimate of -0.552 for Item 2 and fifteen-year-olds indicates that 0.552 must be subtracted from the difficulty of this item for fifteen-year-old students; and the estimates of 0.235 for Item 2 and fourteen-year-olds indicate that 0.235 must be added to the difficulty of this item for fourteen-year-old students. Similarly, the estimate of 0.316 for this item and thirteen-year-olds indicate that 0.316 must be added for thirteen-year-olds. This means that fifteen-year-olds found Item 2 to be easier than the other two groups while thirteen-year-olds found it to be the most difficult.

On the other hand, 21 items were found to be relatively easier for fourteen-year-olds than thirteen- and fifteen-year-olds; while 12 items were easier for thirteen-year-olds than fourteen- and fifteen-year-olds. Fourteen- and fifteen-year-old students also found three items each to be more difficult than the other age cohorts. The remaining 16 items had the same difficulty for test-takers across the three age cohorts. This indicates that at most 74 items would exhibit DIF. However, not all the DIF were of substantive importance.

To determine if the DIF were of substantive importance, the Mantel-Haenszel statistics between each age cohort were calculated. Between the fifteen- and fourteen-year-olds, the magnitude of the Mantel-Haenszel statistics ranged from 0 to 2.841; while those between fifteen- and thirteen-year-olds ranged from 0.02 to 2.691. There were three items with Mantel-Haenszel statistics of above 1.00 for between fifteen- and fourteen-year-olds, two items with Mantel-Haenszel statistics of above 1.00 for

between fifteen- and thirteen-year-olds, and 11 items with Mantel-Haenszel statistics of above 1.00 for between fifteen- and fourteen-year-olds and between fifteen- and thirteen-year-olds. These Mantel-Haenszel statistics however had non-significant chi-square on four degrees of freedom; hence, the sizes of their DIF were considered negligible.

For between thirteen- and fourteen-year-olds, there was only one item with Mantel-Haenszel statistics of above 1.00, i.e. Item 69. With Mantel-Haenszel statistics of 1.772 (p=.068), Item 69 was flagged as having a moderate to large DIF. This means that fourteen-year-olds found the item to be significantly easier than thirteen-year-olds. The DIF contrast was 1.105 logits. It is noteworthy that this is the same item that has a moderate to large DIF across grade levels. Figure 4.12 shows that their empirical curves followed the same trend as those for grade levels; hence, the DIF analysis by age cohorts did not reveal anything new.

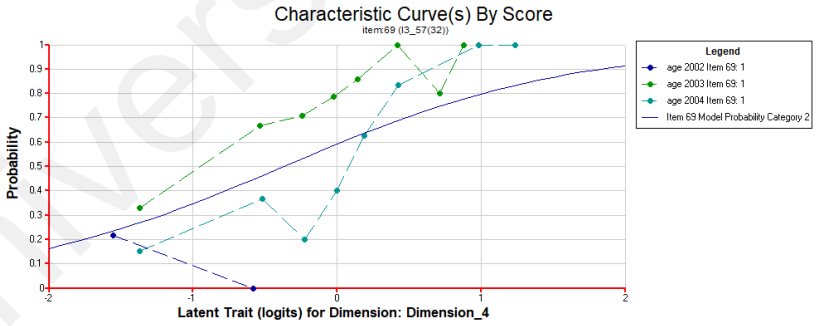


Figure 4.12. Empirical curves for Item 69 by age cohorts.

4.9 Differential Item Functioning Across Genders

Overall, male test-takers have performed more poorly than the females with a parameter estimate of -0.112 (SE=0.006). The actual parameter estimate for the males was more than 18 times larger than its standard error estimate; thus, the mean difference of .224 logits between the male and female students was clearly significant.

The chi-square value of 409.89 on one degree of freedom confirms that the mean difference was statistically significant at $p<.001$. Although the males' mean performance was significantly lower than that of the females, it did not indicate differential item functioning (DIF).

To provide prima facie evidence of DIF, the item parameters estimated from the female subsample were plotted against those from the males. Figure 4.13 shows the cross-plot of item parameter estimates across genders. From the cross-plot, it is apparent that the locations of the measures for Items 1, 9, 11, 38, 43, 48, 57, 60, 64, 67, 71, 81, and 90 varied across genders by more than the modelled errors; while Items 4, 13, 15, 27, 42, 47, 58, 63, 80, 83, and 85 were located at the border of the error band. These items were expected to exhibit DIF. The rest of the items appeared to remain invariant within the modelled errors. The interactions between the item and gender facets were found to be statistically significant, $\chi^2(84) = 1143.01, p < .001$, confirming the existence of DIF in the items.

Appendix J(iii) shows a summary report of the DIF analysis. The report shows that 32 items were relatively easier for males than females; 34 items were relatively easier for females than males; and 24 items had the same difficulty. This means that there were 66 items with different parameter estimates for the male and the female subsamples. Three items that were noticeably off-diagonal in the cross-plot in Figure 4.13 were not included in the list of 66 items, i.e. Items 1, 64, and 90. The estimates of these three items according to gender were either smaller than or equal to their standard errors. Together with the visual inspection from cross-plot, there were at most 69 items that exhibit DIF. The DIF contrast for all the items ranged from 0 to 0.916.

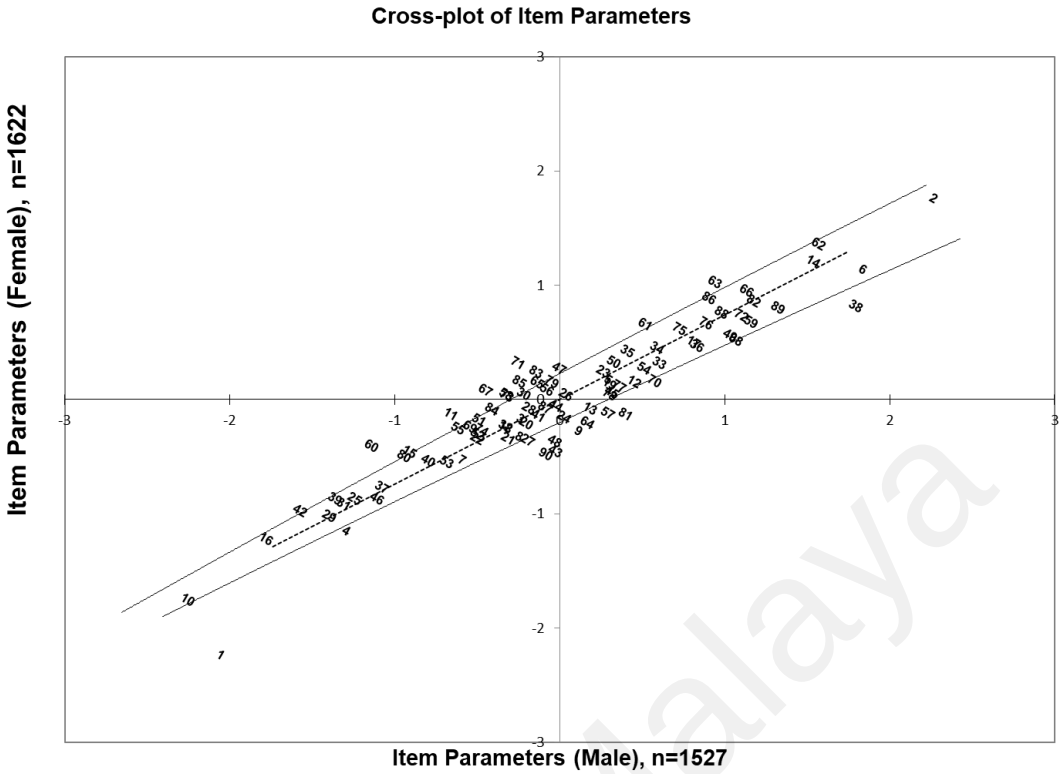


Figure 4.13. Cross-plot of item parameters for the female subsample against the male subsample.

To objectively determine if the DIF in the 69 items were of substantive importance, the Mantel-Haenszel method is used. Except for Items 11, 38, 49, and 60, the Mantel-Haenszel statistics for the remaining 86 items were below 1.00 with non-significant chi-squares on four degrees of freedom, indicating that their DIF were negligible. For Items 11, 38, and 49, the magnitude of their Mantel-Haenszel statistics were below 1.00, but their chi-squares were significant at alpha of .05. This means that the three items had negligible DIF. For Item 60, the Mantel-Haenszel statistic was 1.061 with $\chi^2(4) = 7.342, p = .12$, which was reported as having a slight to moderate DIF. Item 60 was relatively easier for female test-takers than the males by a logit of 0.876. As can be seen in the item characteristic curve in Figure 4.14 below, the empirical curve for females was above that for the males. This suggests that male students found Item 60 to be more difficult than females of comparable ability. Item

60, i.e. Item 60 of Set 1, tested students on their ability to infer meaning of a nonsensical word related to eating behaviours; and thus, there is no apparent logical explanation for the existence of gender DIF for this item.

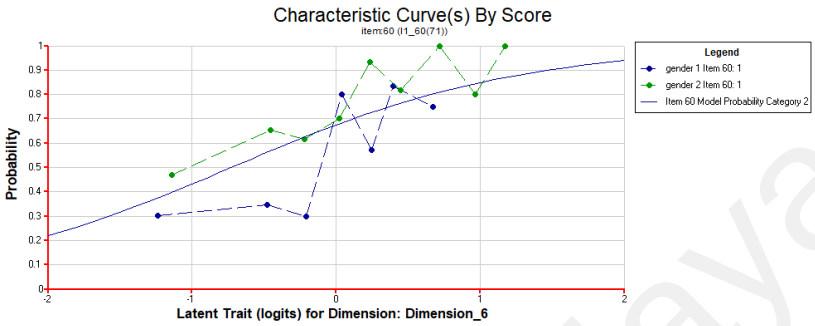


Figure 4.14. Empirical curves for Item 60 by genders.

4.10 Differential Item Functioning Across Ethnic Groups

Students' ethnicity was determined by their parents' ethnicity, so DIF analyses were conducted across their father's and their mother's ethnic groups separately. There were four major ethnic groups in the sample, i.e. Iban, Malay, Chinese, and Bidayuh. Students of other ethnicity such as Orang Ulu were categorised as "Others" for the DIF analyses.

Based on their father's ethnicity, it was found that Iban students outperformed the rest with a parameter estimate of 0.204 (SE=0.010), followed by Malay and Chinese students with parameter estimates of 0.161 (SE=0.010) and 0.093 (SE=0.010) respectively. Similarly, based on their mother's ethnicity, Iban students recorded the best performance in the test with parameter estimate of 0.146 (SE=0.010); while following closely behind were Malay students with parameter estimate of 0.145 (SE=0.010), and Chinese students with parameter estimate of 0.012 (SE=0.011). On the other hand, students of Bidayuh father recorded the poorest test performance with parameter estimate of -0.231 (SE=0.011); while students of Bidayuh mother recorded

the second poorest performance with parameter estimate of -0.084 (SE=0.010). The differences in performance between the ethnic groups were statistically significant, $\chi^2(4) = 1176.39, p < .001$ (based on father's ethnicity), and $\chi^2(4) = 487.74, p < .001$ (based on mother's ethnicity).

The interactions between the item and ethnicity facets were statistically significant: based on father's ethnicity, $\chi^2(336) = 1505.96, p < .001$, and based on mother's ethnicity, $\chi^2(336) = 1458.96, p < .001$. The significant interactions between the facets indicate that DIF exist in the items. Appendix J(iv) shows a summary report of the DIF analyses across ethnic groups. The report shows that six items had the same difficulty for students across ethnic groups based on father's ethnicity, i.e. Items 33, 42, 50, 66, 71, and 87. Meanwhile, seven items had the same difficulty for students across ethnic groups based on mother's ethnicity, i.e. Items 43, 65, 80, 85, 86, 87, and 88. The remaining items were either relatively easier or relatively more difficult for at least one of the ethnic groups. For instance, based on father's ethnicity, Item 52 was relatively easier for Iban students and more difficult for Bidayuh students as compared to the other three groups; but based on mother's ethnicity, it is easier for Malay students and more difficult for Others.

The differences in the difficulty levels across ethnic groups indicate the existence of DIF in the items. However, not all the DIF were of substantive importance. The Mantel-Haenszel statistics together with its chi-square values were computed between each ethnic group to determine if the existence of DIF were of concern. It was found that the Mantel-Haenszel statistics across ethnic groups based on father's ethnicity ranged from 0 to 2.12 in absolute value. There were 14 items with at least one Mantel-Haenszel statistics of more than 1.00, but their chi-square values were not significant. Meanwhile, the magnitude of Mantel-Haenszel statistics across ethnic

groups based on mother's ethnicity ranged from 0 to 2.00, and there were 13 items with at least one Mantel-Haenszel statistics of more than 1.00. However, their chi-square values were not significant. Therefore, none of the items were flagged as having slight to large DIF across ethnic groups. In fact, the suggested DIF category in ConQuest for all the items were A. This means that, even if DIF exist across ethnic groups, they were negligible.

4.11 Differential Item Functioning Across Native Language Clusters

Similar to the DIF analyses across ethnic groups, there were five groupings for DIF analysis across native language clusters, i.e. Iban, Malay, Chinese, Bidayuh, and Others. It was found that Iban speakers outperformed the rest with a parameter estimate of 0.276 (SE=0.10), followed by Chinese and Malay speakers with parameter estimates of 0.210 (SE=0.011) and 0.198 (SE=0.009) respectively. Meanwhile, the parameter estimates for Bidayuh speakers was -0.028 (SE=0.011). Users of other languages recorded the worst test performance with a parameter estimate of -0.656 (SE=0.021). The differences in test performance between the native language clusters were statistically significant, $\chi^2(4) = 1525.03, p < .001$.

The interactions between the item and native language facets were statistically significant, $\chi^2(336) = 1701.71, p < .001$. The significant interactions between the facets indicate that DIF exist in the items. Appendix J(v) shows a summary report of the DIF analyses across native language clusters. The report shows that five items had the same difficulty for students across native language clusters, i.e. Items 50, 60, 66, 85, and 87. The remaining items were either relatively easier or relatively more difficult for at least one of the native language clusters. For instance, Item 63 (Item 57

of Set 2) was relatively easier for Malay speakers but relatively more difficult for other native language users with a DIF contrast of 1.391 between the two groups.

The Mantel-Haenszel statistics for Item 63 between Malay and Others was 3.003, which was the largest in absolute value. However, its chi-square value of 2.114 on four degrees of freedom was not significant (p=0.99). Altogether, there were 17 items with at least one Mantel-Haenszel statistics of greater than 1.00, but their chi-square values were not significant. None of the items were flagged as having slight, moderate, or large DIF across native language clusters. Therefore, the DIF that exists between native language clusters for all the items were negligible.

4.12 Differential Item Functioning Across Geographical Areas

Overall, urban students outperformed rural students by 0.082 logits (SE=0.005). The parameter estimate of 0.041 (SE=.005) for the urban students was eight times larger than its standard error estimate; thus, the difference between the urban and rural means was noticeable. The chi-square value of 56.06 on one degree of freedom confirms that the mean difference was statistically significant at p<.001. In other words, the performance of urban students in the test was significantly better than that of rural students.

Figure 4.15 shows the cross-plot of item parameter estimates across geographical areas. From the cross-plot, it is apparent that the locations of the measures for Items 58, 59, 63, 65, 72, 73, 74, 75, 76, 78, and 79 varied across geographical areas by more than the modelled errors; while Items 1, 28, 56, 57, 60, 81, and 83 were located at the border of the error band. These items were expected to exhibit DIF. The rest of the items appeared to remain invariant within the modelled errors. The interactions between the item and geographical area facets were found to

be statistically significant, $\chi^2(84) = 295.06, p < .001$, confirming the existence of DIF in the items.

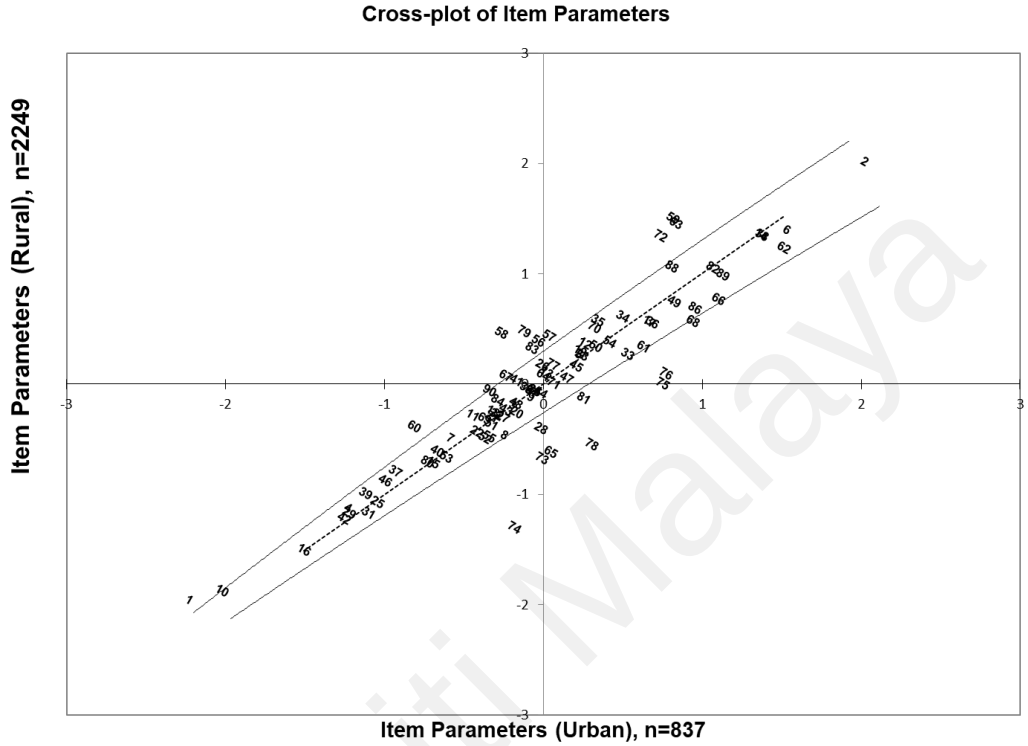


Figure 4.15. Cross-plot of item parameters for the rural subsample against the urban subsample.

Appendix J(vi) shows a summary report of the DIF analysis. The report shows that 23 items were relatively easier for urban students than rural students; 17 items were relatively easier for rural students than urban students; and 50 items had the same difficulty. This means that there were 40 items with different parameter estimates for the urban and the rural subsamples. Item 76 which was noticeably off-diagonal in the cross-plot in Figure 4.15 was not included in the list of 40 items because the item parameter estimates according to geographical areas were smaller than its standard error. Together with the visual inspection from cross-plot, there were at most 41 items that exhibit DIF. The DIF contrast for all the items ranged from 0.004 to 0.856 while

their Mantel-Haenszel statistics ranged from 0.001 to 1.227 in absolute value. Only two items had Mantel-Haenszel statistics of above 1.00, but their chi-squares were not significant on four degrees of freedom. None of the items were flagged as having slight to large DIF. This means that the DIF that exist between urban and rural test-takers were negligible for all the items.

4.13 Summary

The results indicate that the item response data best fit the between-item multidimensionality Rasch model, suggesting that the diagnostic test measures several related unidimensional latent variables. These variables were positively correlated to each other as expected from a multidimensional test of the same construct. However, not all the items and cases fit the model. Out of 90 items and 3,086 cases, the responses of three items did not fit the Rasch model well when the test-takers' abilities were not targeted by the items while 5.83% of the cases underfit the model. The misfits occur most probably due to guessing. There were also 21 items that could not discriminate test-takers of different abilities well. Moreover, it was found that there were gaps in the item distribution across the range of test-takers' abilities for five of the six dimensions although the overall item difficulties were normally distributed. In terms of differential item functioning, one of the items was found to have moderate to large DIF across grade levels and age cohorts while another item had slight to moderate gender DIF. The DIF that exist across ethnic groups, native language clusters and geographical areas were considered negligible. Findings of the study will be discussed in greater depth in the next chapter.

CHAPTER 5

DISCUSSION AND CONCLUSION

Rome was not built in a day.

After trudging through four chapters, Chapter 5 draws the study to a temporary close. The concluding chapter starts with a summary of the findings before discussing them in greater depth. Since the test development is still in its infancy, much work needs to be done; hence, a substantial portion of this chapter is dedicated to a discussion of the different ways the test can be improved upon. This would provide some direction for further studies into the test. Implications from the findings and recommendation for future research conclude the chapter.

5.1 Summary of the Findings

The goal of the diagnostic test was to identify the strengths and weaknesses of individual students' linguistic competence. Every item in the test was written by drawing upon the Form 1 and Form 2 English language syllabi and has undergone judgement from seven experts with diverse experiences. The items were then modified, selected, and assembled into six sets with 54 common items and six non-link items each. Altogether, there were 90 items divided into six domains: graphology, lexical items, word classes, morphology, syntax, and semantics. The six sets were then administered at random to a representative sample of 3,086 students of lower secondary schools in the southern zone of Sarawak.

The item response data were initially fitted to the unidimensional Rasch model using Winsteps, but the results suggest possible multidimensionality. Thus, the data were fitted to the unidimensional, subdimension, and between-item multidimensionality Rasch models using ConQuest. Among the three variants of the Rasch model, it was found that the item response data best fit the between-item six-dimensional model. This strongly suggests that the test measures several related unidimensional latent variables. To be specific, the dimensionality analyses show that the diagnostic test of linguistic competence measures six positively correlated dimensions.

Although the item response data best fit the multidimensionality Rasch model, not all the items and cases fit the model. The unweighted mean-squares and item characteristic curves indicate that the responses to Items 38, 59 (59 of Set 1), and 89 (59 of Set 6) did not fit the model well when the test-takers' abilities were much lower than those targeted by the items. These three were the most difficult items within the syntax dimension. Similarly, the person fit statistics and kidmaps indicate that 5.83% of the cases underfit the model; however, their presence did not influence the item calibration. This suggests that the item response data fit the Rasch model reasonably well; and any misfits that occur are minimal and most probably due to guessing.

Meanwhile, the scatterplot of the unweighted mean-squares against the point-biserial correlation discrimination estimates shows that all the 21 items with point-biserial below the recommended guideline of .20 had unweighted mean squares larger than the expected mean square of 1.00, indicating that less discriminating items did not fit the Rasch model very well. Out of these 21 items, 17 items also had at least one distractor with positive point-biserial correlation. However, for the 69 items with point-biserial discrimination above .20, 41 items had mean squares of 1.00 and below

while 28 items had mean squares above 1.00. This seems to suggest that most of the items had point-biserial discrimination between .20 and .57; thus, items with discrimination outside this range would not fit the Rasch model well. In other words, the unequal item discrimination could be a factor behind the misfits for the less discriminating items. For items that can discriminate test-takers of different abilities relatively well but did not fit the Rasch model, other factors such as guessing might be contributing to the misfits. As such, there is a need to further investigate some of the items to determine if they can be kept in the item bank for future use.

Despite the misfits and less discriminating items, the reliability indices show that the relative locations of the Rasch measures for both the items and persons were reproducible. Specifically, the sample of test-takers was sufficiently large and diverse to confirm the hierarchy of item difficulties; and the range of items for the dimensions of lexical items, word classes, and syntax was sensitive enough to distinguish test-takers of different abilities. However, the range of items for the dimensions of graphology, morphology, and semantics might not be sensitive enough to distinguish between low and high achievers.

The Wright map confirms that there were gaps in the item distribution across the range of test-takers' abilities, especially for the dimensions of graphology, morphology, and semantics. For the graphology dimension, there were big gaps at the lower end and upper end of the logit scale; while for morphology and semantics, there were no items at the lower end of the scale, confirming that the range of items was not sensitive enough to differentiate low achievers from high achievers. In contrast, for the syntax and word classes dimensions, the items were concentrated in the middle of the scale, and thus, were able to distinguish between low and high achievers. As for lexical items, the items covered the range of test-takers with very little gaps in between.

Despite the gaps in the item distribution per dimension, the items in general were reasonably well-matched to the test-takers' abilities.

Finally, the differential item functioning analyses indicate that there were very few items that function differently across the different demographic groups even though there were significant differences in test performance between the groups. Across grade levels, only Item 69 (57 of Set 3) exhibits a moderate to large DIF that favours Form 2 students. The same item also displays a moderate to large DIF between thirteen- and fourteen-year-olds, suggesting that the item was indeed easier for Form 2 students and fourteen-year-olds. Another item, Item 60 (60 of Set 1), exhibits a slight to moderate gender DIF that favours females. Across ethnic groups, native language clusters, and geographical areas, the DIF that exist were not substantive enough to be of concern. In other words, except Items 60 and 69, the measurement remains invariant across the different demographic contexts.

5.2 Discussion of the Findings

5.2.1 Investigation of items with low point-biserial discrimination estimates

Findings of the study show that there were 21 items with low point-biserial correlation (see Appendix I). At first glance, these items were problematic because they were not able to discriminate between test-takers of different abilities. Specifically, test-takers who failed to respond to the items correctly tend to perform well in the test overall while those who answered the items correctly tend to perform poorly in the overall test (Coaley, 2010; Finch, Immekus, & French, 2016; Osterlind, 2006; Varma, 2010). This is an anomaly; hence, the 21 items need further investigation. Out of the 21 items, three belong to the graphology domain, three lexical items, two

testing word classes, eight measuring morphological knowledge, none testing syntax, and five in the semantics dimension.

It is interesting to note that the three items in the graphology domain were the three most difficult items within the dimension. For Item 2 (r_{pb}=.08) where students had to choose the correct spelling for a visual stimulus, distractor C (*stationary*) with point-biserial of .16 was a competing answer option that appealed to most of the test-takers although their dimensional mean ability of 0.41 was lower than the mean ability of 0.61 for those who opted for the correct answer D (*stationery*). The same trend is also observed for Item 6 (r_{pb}=.15) which tested students' ability to use the apostrophe s. The popular choice for distractor C in both items perhaps can be explained by the concept of middle bias (Attali & Bar-Hillel, 2003), where examinees who guess the answer have a strong and systematic tendency to seek it in the middle position. Students who were confused between options C and D most probably had selected C because it was in the middle position. To reduce the effect of the middle bias, these two items can be revised by repositioning option C at the edge (i.e. in the A position). For Item 62 which required students to identify the correctly spelt word without any context, 26% and 36% of test-takers selected distractors B (*acommodate*) and C (*accomodate*) respectively although their mean ability was lower than the 26% who opted for the correct answer A (*accommodate*). Perhaps, Item 62 can be revised by changing the item stem to include the context where the word *accommodate* is used.

Similar to the graphology dimension, the three least discriminating lexical items were also the three topmost difficult items within its domain. For Item 14 on expression of opinions, distractors B (*is according to me*) and C (*believes largely*) were compelling alternatives to the answer key A (*goes without saying*). Although A is the best answer option, B and C are possible answers to the stem (*It* _____ *that computers*

help us in many ways). Option B is grammatically correct despite being an awkward expression in the context of this item while option C, if rephrased as *is largely believed*, can be an answer. Since the answer options cannot be easily fixed, it is perhaps better to replace Item 14. For Items 63 and 82, despite the low item point-biserial discrimination, none of the distractors had positive point-biserial. This suggests that Items 63 and 82 are relatively less discriminating but not faulty. Therefore, these two items can be retained without much modification.

Within the word classes dimension, Item 88 with point-biserial correlation of .17 was the most difficult item. Although the item point-biserial discrimination was relatively low, the point-biserial correlation for all the distractors were negative, suggesting that there were no competing alternative answers. Therefore, Item 88 can be retained without modification. For Item 23 with point-biserial discrimination of .14, the point-biserial correlation for one of its distractors was positive but low (r_{pb}=.01). Furthermore, none of the distractors appeared to be an alternative answer during the expert judgement stage (see Appendix F). Since the distractor did not appear to be a compelling alternative, it is suggested that Item 23 be retained without modification.

Out of the eight items with low point-biserial discrimination in the morphology dimension, only Item 32 had negative point-biserial for all the distractors, suggesting that there were no competing answer options; thus Item 32 can be kept. Apart from Item 32, the remaining seven items were the second to the eighth topmost difficult items in the morphology dimension. It must be noted that the most difficult item was Item 86 with an item point-biserial discrimination of .35 and a positive but lower point-biserial correlation for one of its distractors (r_{pb}=.13). Because the mean ability for the answer key was higher than the mean ability for the distractor, it can be concluded that Item 86 was able to discriminate test-takers at the upper end of the ability continuum.

Similar argument can also be made for the second most difficult item in the morphology dimension Item 36, which had point-biserial discrimination of .12 and positive but much lower point-biserial correlation for distractors B (r_{pb}=.02) and D (r_{pb}=.04), indicating that distractors B and D were attracting students of higher ability but it was the answer key that attracted the most competent students. Therefore, Item 36 can be retained.

Unfortunately, the remaining five morphological items with low point-biserial discrimination had at least one competing alternative answer. For example, Item 34 with point-biserial discrimination of .05 had distractor A with a higher point-biserial of .07 while Items 35 and 64 with point-biserial discrimination of 0 had distractor D with point-biserial of .12 and .20 respectively. For Item 33, the point-biserial of .14 for its distractor B was almost equal to its item point-biserial of .15 while for Item 77, the point-biserial of 0 and .03 for its distractors C and B respectively were rather close to its item point-biserial of .07. As it is not easy to rectify the distractors for these items, it is suggested that these items be removed from future test.

Out of the five semantic items with low point-biserial discrimination, only Item 48 can be retained in its original form. This is because the point-biserial of .05 for its distractor D was much lower than its item point-biserial of .19, indicating that there was no competing alternative answer. It is worth noting that the remaining four least discriminating items were the four topmost difficult items within the semantic dimension. For Item 49 with point-biserial discrimination of .13, which required students to select the best definition for a word used in context, distractor C had a higher positive point-biserial correlation (r_{pb}=.16) but lower mean ability than the answer key. To improve the item point-biserial discrimination, distractor C can be replaced with similar but simpler definition so that test-takers who knew the meaning

of the word in the stem can better comprehend the definition in the distractor. In contrast, for Items 66 and 72, one of the distractors attracted students with higher mean ability as compared to the answer keys, indicating that the competing distractors could have qualified as answers. These two items also had extremely low item point-biserial correlation. Another item, Item 54 which required students to detect anomalous sentences, had an item point-biserial discrimination of .05 and point-biserial of 0 and .05 for distractors A and D respectively. Therefore, Items 54, 66, and 72 cannot discriminate test-takers of different abilities and thus should be dropped from the test.

From the investigation of items with low item point-biserial discrimination, it is pertinent to note that the recommendation of having items with point-biserial discrimination of at least .20 (Shultz, Whitney, and Zickar, 2014) and negative point-biserial correlation for each distractor (Millman & Green, 1989) can sometimes be argued against. This is especially the case when the point-biserial correlation is interpreted within the framework of Rasch model as were done in the current study. For instance, when the item difficulty and the mean ability of test-takers for each distractor are taken into consideration, the less-than-.20 point-biserial discrimination and the positive point-biserial correlation for the distractor often appear to be acceptable. In fact, Wright (1992) argues that the range of acceptable or desirable point-biserial correlation based on raw scores are unknown. As such, the recommended range of point-biserial correlation under classical test theory should serve only as a guide and not an absolute cut-off value. However, test developers should not ignore the item point-biserial discrimination when investigating the quality of the test. In fact, the point-biserial correlation are rather useful as a complement to Rasch analysis, especially for items in the multiple-choice format. Moreover, the scatterplot of fit statistics against item point-biserial discrimination can provide a clue

as to why some items do not fit the Rasch model as well as the others. In this study, it has been shown that the heterogeneity of the least discriminating items might explain the less-than-ideal fit statistics for these items.

5.2.2 Data-model fit

Findings of the study show that three items misfit the six-dimensional Rasch model in terms of unweighted mean squares, indicating that the responses of test-takers who were not targeted by these items were more erratic than expected by the Rasch model. It is to be noted that the underfitting items were the three topmost difficult items within the syntax dimension. When students with low syntactical competence encountered these items, they might have resorted to guessing. Because the Rasch model only models two aspects of a testing situation – item difficulty and person ability, any extraneous factor is considered residuals and counted towards misfit (Bond & Fox, 2015). Because the underfitting items had item point-biserial discrimination that were within those of the majority of the items, unequal item discrimination can be ruled out as a possible factor in this case. Therefore, guessing is the most probable reason behind the misfit, which is expected in multiple-choice items. As such, the three underfitting items would not be removed from the test.

It also must be noted that there is no strict rule to the acceptable range of fit statistics; and thus, the assessment of fit is a balancing act and a judgement call (Bond & Fox, 2015; Douglas, 1982; Wu et al., 2016). For instance, the *t* statistic is sensitive to sample size. Since the sample size of this study was large (N=3,086), it was not suitable to assess the data-model fit based on the *t* statistic. Instead, the weighted and unweighted mean squares were used in the fit assessment. Because the test in this study was diagnostic in nature as opposed to a high-stake test, the range of acceptable mean

square was set at 0.50 and 1.00 in the Bond-and-Fox developmental pathway. This range is just a guideline in the fit assessment; and thus, it must be clarified that all the items that were considered to fit the Rasch model in this study did not fit the model perfectly but were only within the acceptable range.

Findings of the study also demonstrate that the Rasch requirement for unidimensionality did not hold up empirically in the data, and that the data were best fitted to the Rasch between-item multidimensionality model. The multidimensionality in the item response data is somewhat anticipated as the test was designed from the outset to measure six domains of linguistic competence. Each domain is unidimensional on its own and are significantly correlated with each other. The multidimensional model allows collateral information from other domains to be drawn upon when estimating the test-taker's ability in one domain (Wu, Tam, & Jen, 2016). This is more advantageous than the consecutive approach, which is simply a unidimensional Rasch model being repeated separately for each domain, because it enhances reliability and more accurately represents students' performance (Briggs & Wilson, 2003).

By applying the multidimensional Rasch model, test-taker's ability estimate in each domain is determined not only by the raw score in the domain but also by the raw scores in other domains. For example, both Persons 1919 and 666 responded to seven out of ten graphology items correctly, but the ability estimate in the graphology dimension for Person 1919 was 0.87 logit as compared to 1.11 logit for Person 666. Because the graphology domain correlates with the other dimensions, and Persons 1919 and 666 had different raw scores in the other dimensions, the multidimensional Rasch model provides different estimates of their graphological ability.

In this study, the item response data were also fitted to the Rasch unidimensional and subdimension models, which were found to have poorer fit than the multidimensional model. However, it must be noted that data-model fit is not absolute and depends on the pool of items in the test (Wu, Tam, & Jen, 2016). This means that, if the test were to be revised, the item response data collected might fit the subdimension or the unidimensional model better. In the revised test, problematic items would have to be removed or modified. Items would also have to be added to fill the gaps in the Wright map (see Figure 4.9) so that the test-takers' abilities can be better targeted by the items. For example, the revised test needs easier morphology and semantic items. Item response data to be collected using the revised test may fit the models differently since fit statistics are relative.

5.2.3 Test usefulness

The multidimensional Rasch model estimates the ability of each test-taker in every one of the six dimensions without estimating any composite or overall ability. For example, Person 666 was estimated to have an ability of 1.11 logit in graphology, 0.34 logit in lexicon, 0.69 logit in word classes, -0.91 logit in morphology, 1.79 logit in syntax, and -0.39 logit in semantics. As a diagnostic test, the separate ability estimates are sufficient for test users to make decision regarding test-takers' areas of strengths and weaknesses. For example, if Person 666 were to improve her linguistic competence, the focus should be on morphology and semantics. An overall ability estimate of her linguistic competence is not necessary in this case. Another useful diagnostic tool is the kidmap of the test-taker. This is because a close inspection of the individual kidmap can reveal unexpected patterns of responses that are always worth investigating if diagnosis is the intent of testing. The kidmap might also explain why

the test-takers misfit the Rasch model. Although the test is a diagnostic tool, it is the test users that initiate the diagnosis process (Alderson et al., 2015).

Teachers can make use of information from the Rasch analyses to diagnose students' strengths and weaknesses in linguistic competence. The diagnostic feedback can be linked to intervention plans to help students improve their areas of weaknesses. Since the test specifications are drafted based on the Form 1 and Form 2 syllabi, teachers can design their lessons directly based on the test results. For instance, if the teacher found that most students in the class has gaps in their semantical knowledge, the teacher might want to spend a few lessons in this area. However, if only a few students have problems with their semantical competence, the teacher might want to design individualised tasks to help these few students. The test developed in this study has the potential to offer diagnostic information that can help teachers to identify specific areas of linguistic competence that need attention, hence making it a relevant and useful diagnostic tool (Alderson et al., 2015; Brown & Abeywickrama, 2010; Lee, 2015).

It is not difficult for teachers to score the test since all the items are in the multiple-choice format. What is difficult however is for teachers to analyse the scores within the framework of Rasch model and interpret the results. To make the process simpler for teachers, multidimensional computerised adaptive testing can be applied to the diagnostic test. The current study provides an item pool that has been calibrated on a representative sample of targeted test-takers. It also provides some direction on the types of items to be added and how existing items can be modified. With new online calibration methods for multidimensional computerised adaptive testing, it would not be too difficult to replenish the item bank in the future (Chen, Wang, Xin, & Chang, 2017).

Multidimensional adaptive testing provides a highly efficient approach to measure test-takers' abilities in several latent traits by incorporating information from several sources on all dimensions simultaneously (Frey, Seitz, & Brandt, 2016). For example, a correct response to a morphology item would increase the provisional estimates not only for the morphology dimension but also for the other five dimensions but at different rates. This is because all the six dimensions are positively correlated at different magnitudes. The added information from items of other correlated dimensions can lead to reduced test-lengths and greater measurement precision. Moreover, multidimensional adaptive testing ensures adequate content coverage at appropriate difficulty level by using information from prior joint-distribution of ability in its item selection algorithm (Segall, 1996). With multidimensional adaptive testing, the usefulness of the diagnostic test would be enhanced.

5.2.4 Test development

The possibility of applying multidimensional adaptive testing to the current diagnostic test indicates that the test development process has not ended. In fact, the pencil-and-paper version of the diagnostic test developed in this study is just the beginning. This reflects the iterative nature of test development (Markus & Borsboom, 2013). Much work still needs to be done if the diagnostic test were to provide reasonably precise measurement of test-takers' linguistic competence in a user-friendly manner.

First, the underfitting items, items with problematic distractors and items with substantive DIF must be fixed, removed, or noted. For example, Item 69 which exhibits moderate to large DIF that favours Form 2 students can either be removed from the test or only administered to Form 2 students so that it does not disadvantage

the Form 1 students. Next, more items need to be drafted to fill the gaps in item difficulty, especially in graphology, morphology, and semantics dimensions. This will increase the sensitivity of the item range so that low achievers can be distinguished from high achievers. For word classes and syntax dimensions, easy items as well as difficult items are needed so that test-takers at both ends of the scale can be better targeted. Even when the gaps have been filled, new items still need to be drafted to replenish the item bank to ensure test security. These new items must be calibrated using a representative sample of test-takers before they are added to the test.

Item calibration using the between-item multidimensionality Rasch model is just one of the many options that test developers could use. A less computationally demanding alternative is to fit the items within each dimension to unidimensional Rasch model separately using the consecutive approach (Linacre, 2009). However, with just 15 items per dimension, separate unidimensional analyses would greatly reduce the amount of information for item calibration and measurement of persons, resulting in less precision. Another alternative is to model the item response data as a function of diagnostic states on categorical latent variables using a diagnostic classification model (Rupp, Templin, & Henson, 2010). This approach classifies test-takers statistically into mastery decisions and can provide a finer grain size analysis, which are often desired in diagnostic reporting. However, specifying the attribute structure and estimating the diagnostic classification model are complex processes that are beyond the scope of the current study.

5.3 Implications of the Study

It is clear from the findings of the study that the diagnostic test of linguistic competence in the English language has strong potential. As it is, the test can be used to measure lower secondary school students' linguistic competence with reasonable precision. This has significant implications in the field of language teaching in Malaysia. Since the test is aligned with the lower secondary school syllabi, the test can be incorporated directly into any language programme that uses the Malaysian secondary school curriculum. For instance, the test can be administered to students on the first week of secondary school to help English language teachers identify the gaps in their students' linguistic competence. Using results from the analyses of students' responses, teachers can develop their lessons and/or design individualised tasks to help students overcome their weaknesses. This means that the diagnostic test is an assessment that can promote learning.

The study has also demonstrated that it is possible to develop an English language test that is purely diagnostic, which has a serious implication for the language testing industry. At present, there is a scarcity of true diagnostic assessment in second and foreign language (Alderson et al., 2015), hence leaving a void in the language testing industry waiting to be filled. The test development process described in the current study can be replicated to create more localised diagnostic language tests. For example, similar procedures can be used to develop diagnostic tests for the Malay language and the Chinese language. It must, however, be noted that the current test development process is just the beginning. The current pencil-and-paper test provides well-calibrated items that can be deposited into the item bank of a multidimensional computerised adaptive test. Computerised adaptive testing is a game changer in the field of language testing because of its capability to provide instantaneous scoring and

reporting. With advanced computing technology, the testing system can be integrated with a self-directed learning system. For example, a student reported to have a low ability estimate in the dimension of syntax can be directed to learning materials in the domain that are targeted at his ability.

Furthermore, findings of the study have shown that the item response data best fit a multidimensional model. This has some indirect implications for the development of language theories. First, the multidimensionality that exists in the data implies that the construct being measured might also be multidimensional in nature. This means that empirical data have verified that linguistic competence is comprised of multiple latent variables that are positively correlated to each other. It is only possible to build a test of linguistic competence because it has been well recognised in the literature that linguistic competence is comprised of several dimensions such as morphology, syntax, and semantics. This implies that language ability must be well theorised so that items can be written directly to measure the ability. The study therefore provides an example where theory drives the test construction and empirical data collected from the test verify the theory.

5.4 Recommendations for Future Research

Despite its promising potential, the current study is admittedly a work in progress. There are various directions on which the current work can be further advanced. First, the diagnostic English language test developed in this study only measures linguistic competence. There are other components in the models of communicative competence such as discourse competence, sociolinguistic competence, and strategic competence. Therefore, the next step is to draft and calibrate items to measure these components so that a complete test of communicative

competence can be made available to teachers who are interested in diagnosing the language needs of their students.

It is also important to study the impact of the diagnostic English language test on real classroom teaching and learning. This involves obtaining evidence on actual consequences of how the test results are interpreted and used by the stakeholders, especially teachers, students, parents, and school administrators. It would defeat the purpose of diagnostic testing if the test does not lead to improvement in students' learning. For instance, it would be an additional burden for teachers if the test scores were merely reported and not used to help overcome students' weaknesses in linguistic competence. Moreover, it would be detrimental to the teaching and learning process if the test results lead to unintended consequences such as being used as a tool to evaluate teacher or school performance.

Another direction that the study could have been expanded upon is to computerise the diagnostic test using multidimensional adaptive testing. Specifically, the items calibrated on a representative sample of lower secondary school students in the current study can be deposited into the item bank for the adaptive test. Using the dataset from the current study, simulations can be conducted to select an optimal estimation method, item selection rule and stopping rule. After designing the testing system with a user-friendly interface, the computerised diagnostic test is ready to be administered to test-takers, and this will start off the next phase of test development.

5.5 Conclusion

Findings of the study suggest that the diagnostic test can pinpoint students' weaknesses in linguistic competence, and thus can help teachers tailor their teaching according to students' language needs. However, the test is still in its infancy with

several directions that future research can embark on. Specifically, it has a promising potential to be developed into a multidimensional computerised adaptive test. This has serious implications for the language testing industry. So long as investment flows in, the test can always be further improved, and thus test development is a never-ending process. Another possibility is for the test to be integrated into English language lessons either in real classroom or in a virtual environment. This would be an ideal world where testing, teaching, and learning are well aligned. Much work still needs to be done just like the building of Rome; hence, this chapter only draws a temporary close to the study.

REFERENCES

Abdul Rashid, M., Lin, S. E., & Shaik Abdul Malik, M. I. (2012). The potency of 'READS' to inform students' reading ability. *RELC Journal, 43*(2), 271-282. doi:10.1177/0033688212451803

Adams, R. J. (2010). *Case (person) fit and residuals.* Retrieved from https://www.acer.org/files/Conquest-Notes-3-CaseFitAndResiduals.pdf

Adams, R. J., & Wu, M. (2010). *ConQuest tutorial 7: Multidimensional models*. Retrieved from https://www.acer.org/files/Conquest-Tutorial7.zip

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57-76). New York, NY: Springer Science.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R., Wu, M., Macaskill, G., Haldane, S., & Sun, X. X. (2017). ConQuest (Version 4.14.2) [Computer software]. Victoria, Australia: Australian Council for Educational Research.

Alderson, J. C. (1981). Report of the discussion on testing English for specific purposes. In J. C. Alderson, & A. Hughes (Eds.), *Issues in language testing* (pp. 123-134). London, UK: British Council.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* New York, NY: Continuum.

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 21-39). Ontario, Canada: University of Ottawa Press.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236-260. doi:10.1093/applin/amt046

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge, UK: Cambridge University Press.

Alhadjri, A. (2017, May 25). Some secondary school students still illiterate in English. *Malaysiakini*. Retrieved from http://www.malaysiakini.com/

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), Supplement.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Arabski, J., & Wojtaszek, A. (2011). Introduction. In J. Arabski, & A. Wojtaszek (Eds.), *Aspects of culture in second language acquisition and foreign language learning* (pp. 1-4). Berlin, Germany: Springer-Verlag.

Arukesamy, K. (2015, May 17). Students losing out due to lack of English skill. The Sun Daily. Retrieved from http://www.thesundaily.my/

Ary, D., Jacobs, L. C., & Razavieh, A. (2002). *Introduction to research in education* (6th ed.). Belmont, CA: Wadsworth.

Aryadoust, S. V. (2009). Mapping Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions, 23*(1), 1192-1193.

Asmah Haji Omar. (1982). *Language and society in Malaysia.* Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109-128.

Aubrey, S. (2017a, May 21). Towards English proficiency. *Borneo Post Online*. Retrieved from http://www.theborneopost.com/

Aubrey, S. (2017b, May 30). Taking the first step. *Borneo Post Online*. Retrieved from http://www.theborneopost.com/

Aubrey, S. (2017c, May 21). Towards English proficiency. *Borneo Post Online*. Retrieved from http://www.theborneopost.com/

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bachman, L., & Palmer, A. (1984). Some comments on the terminology of language testing. In C. Rivera (Ed.), *Communicative competence approaches to language proficiency assessment: Research and application* (pp. 34-43). Avon, England: Multilingual Matters.

Bagarić, V., & Djigunović, J. M., (2007). Defining communicative competence. *Metodika, 8*(1), 94-103.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10*(1), 233-252.

Baker, F. B. (1998). The Basics of Item Response Theory (Version 1.0) [Computer Software]. Retrieved from http://ericae.net/irt/baker/

Bell, R. C. (1982). Person fit and person reliability. *Education Research and Perspectives, 9*(1), 105-113.

Bogue, E. L., DeThorne, L. S., & Schaefer, B. A. (2014). A psychometric analysis of childhood vocabulary tests. *Contemporary Issues in Communication Science and Disorders, 41*, 55-69.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier, & D. Hastedt (Eds.), *IERI monograph series: Vol. 1. Issues and methodologies in large-scale assessments* (pp. 51-70). Hamburg, Germany: IEA-ETS Research Institute.

Brandt, S. (2017). Concurrent unidimensional and multidimensional calibration within item response theory. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana, 54*(2), 1-18.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Brown, H. D. (2014). *Principles of language learning and teaching: A course in second language acquisition* (6th ed.). White Plains, NY: Pearson Education.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R., & Courtis, S.A. (1921). Report of the standardization committee. *Journal of Educational Research, 4*(1), 78-80.

Cambridge English. (2013). *Results report: Cambridge baseline 2013*. Cambridge, UK: Author.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In C. Richards, & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London, UK: Longman.

Canale, M. (1984). A communicative approach to language proficiency assessment in a minority setting. In C. Rivera (Ed.), *Communicative competence approaches to language proficiency assessment: Research and application* (pp. 107-122). Avon, England: Multilingual Matters.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Canale, M., & Swain, M. (1981). A theoretical framework for communicative competence. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence.* Washington, DC: Teachers of English to Speakers of Other Languages.

Carroll, J. B. (1961). *Testing the English proficiency of foreign students.* Washington, DC: Center for Applied Linguistics.

Carson, K., Boustead, T., & Gillon, G. (2014). Predicting reading outcomes in the classroom using a computer-based phonological awareness screening and monitoring assessment (Com-PASMA). *International Journal of Speech-Language Pathology, 16*(6), 552-561. doi: 10.3109/17549507.2013.855261

Cazden, C. B. (1996, March 23-26). *Communicative competence: 1966-1996*. Paper presented at the Annual Meeting of the American Association for Applied Linguistics,
Chicago, IL.

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. A. Soler, & M. P. S. Jordà (Eds.), *Intercultural language use and language learning* (pp. 41-57). Castelló, Spain: Springer.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). A pedagogical framework for communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5-35.

Ch'ng, L.-C., & Rethinasamy, S. (2013). English language assessment in Malaysia: Teachers' practices in test preparation. *Issues in Language Studies, 2*(2), 24-39.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385-405. doi:10.1177/0265532214565386

Chen, P., Wang, C., Xin, T., & Chang, H.-H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 70*(1), 81-117.

Chia, J. (2017, August 7). New initiatives to help improve English language proficiency in schools. *Borneo Post Online*. Retrieved from http://www.theborneopost.com/

Chomsky, N. (1965). *Aspects of the theory of syntax*. Massachusetts, MA: Massachusetts Institute of Technology.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods, 17*(1), 31-43.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics.* London, UK: Sage.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Council of Europe. (2011). *Manual for language test development and examining: For use with the CEFR*. Strasbourg, France: Author.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: Sage Publications.

Cronbach, L. J., & Azuma, H. (1962). Internal-consistency reliability formulas applied to randomly sampled single-factor tests: An empirical comparison. *Educational and Psychological Measurement, 22*, 645-665.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cumming, A. (2015). Design in four diagnostic language assessments. *Language Testing, 32*(3), 407-416. doi:10.1177/0265532214559115

DeMars, C. (2010). *Item response theory.* New York, NY: Oxford University Press.

Deming, W. E. (1960). *Sample design in business research.* New York, NY: Wiley.

Douglas, G. (1982). Issues in the fit of data to psychometric models. *Education Research and Perspectives, 9*(1), 32-43.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt am Main, Germany: Peter Lang.

Education First. (2011). *EF EPI English proficiency index*. Retrieved from http://www.ef.com/epi

Education First. (2012). *EF EPI English proficiency index*. Retrieved from http://www.ef.com/epi

Education First. (2013). *EF EPI English proficiency index*. Retrieved from http://www.ef.com/epi

Education First. (2014). *EF EPI English proficiency index*. Retrieved from http://www.ef.com/epi

Education First. (2015). *EF EPI English proficiency index*. Retrieved from http://www.ef.com/epi

Education First. (2016). *EF EPI English proficiency index*. Retrieved from http://www.ef.com/epi

English Language Standards and Quality Council. (2015). *English language education reform in Malaysia: The roadmap 2015-2025*. Putrajaya, Malaysia: Ministry of Education Malaysia.

Facon, B., & Magis, D. (2016). An item analysis of the French version of the test for reception of grammar among children and adolescents with Down syndrome or intellectual disability of undifferentiated etiology. *Journal of Speech, Language, and Hearing Research, 59*, 1190-1197. doi:10.1044/2016_JSLHR-L-15-0179

Finch, W. H., Immekus, J. C., & French, B. F. (2016). *Applied psychometrics using SPSS and AMOS*. Charlotte, NC: Information Age Publishing Inc.

Fishman, J. A., & Galguera, T. (2003). *Introduction to test construction in the social and behavioral sciences*. Lanham, MD: Rowman & Littlefield Publishers.

Fletcher, J., Hogben, J., Neilson, R., Lalara, R. D., & Reid, C. (2015). Examining the quality of phonological representations in Anindilyakwa children in Australia. *International Journal of Language & Communication Disorders, 50*(6), 842-848. doi: 10.1111/1460-6984.12174

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York, NY: McGraw-Hill.

Frey, A., Seitz, N.-N., & Brandt, S. (2016). Testlet-based multidimensional adaptive testing. *Frontiers in Psychology, 7*, 1758. doi:10.3389/fpsyg.2016.01758

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77-92. doi:10.1177/0265659009349972

Fridman, A. E. (2012). The quality of measurements: A metrological reference (A. Sabak & P. Makinen, Trans.). New York, NY: Springer.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Oxon, OX: Routledge.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123-144. doi:10.1177/0265532208097339

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246.

Goldstein, H., & Blinkhorn, S. (1982). The Rasch model still does not fit. *British Educational Research Journal, 8*(2), 167-170.

Granfeldt, J., & Ågren, M. (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing, 31*(3), 285-305.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17*(2), 16-19.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317-336. doi:10.1177/0265532214564505

Hoffman, L. M., Loeb, D. F., Brandel, J., & Gilliam, R. B. (2011). Concurrent and construct validity of oral language measures with school-age children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54,* 1597-1608. doi:10.1044/1092-4388(2011/10-0213)

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60,* 523-531.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

IELTS. (2015). *IELTS performance for test takers 2015.* Retrieved from https://www.ielts.org/teaching-and-research/test-taker-performance

Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Katzenberger, I., & Meilijson, S. (2014). Hebrew language assessment measure for preschool children: A comparison between typically developing children and children with specific language impairment. *Language Testing, 31*(1), 19-38. doi:10.1177/0265532213491961

Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227-258. doi:10.1177/0265532214558457

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509-541. doi:10.1177/0265532211400860

Kostelecká, Y., Kostelecký, T., Vodičková, K., & Jančařík, A. (2015). Linguistic integration of middle school immigrant children in Czechia. *AUC Geographica, 50*(2), 181-192. doi:10.14712/23361980.2015.97

Kumar, R. (2011). *Research methodology: A step-by-step guide for beginners* (3rd ed.). London, UK: Sage.

Kunnan, A. J. (2008). Large scale language assessments. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 135-155). New York, NY: Springer.

Lado, R. (1961). *Language testing.* New York, NY: McGraw-Hill.

Language Testing Research Centre. (2009). *The diagnostic English language assessment (DELA): Handbook for candidates.* Melbourne, Australia: The University of Melbourne.

Le, L. (2012). *Item point-biserial discrimination.* Retrieved from https://www.acer.org/files/Conquest-Notes-5-ItemPointBiserialDiscrimination.pdf

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*(3), 229-316. doi:10.1177/0265532214565387

Lembaga Peperiksaan. (2015). *Pengumuman analisis keputusan sijil pelajaran Malaysia (SPM) tahun 2015* [Announcement of the analysis of the Malaysia education certificate (SPM) year 2015 results]. Putrajaya, Malaysia: Kementerian Pendidikan Malaysia.

Lembaga Peperiksaan. (2016). *Pengumuman analisis keputusan ujian pencapaian sekolah rendah (UPSR) tahun 2016* [Announcement of the analysis of the primary school achievement test (UPSR) year 2016 results]. Putrajaya, Malaysia: Kementerian Pendidikan Malaysia.

Letts, C., Edwards, S., Schaefer, B., & Sinka, I. (2014). The new Reynell developmental language scales: Descriptive account and illustrative case study. *Child Language Teaching and Therapy, 30*(1), 103-116. doi:10.1177/0265659013492784

Lewis, M. (2002). *The lexical approach: The state of ELT and a way forward.* Boston, MA: Thomson.

Li, H., & Suen, H. K. (2012). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30*(2), 273-298. doi:10.1177/0265532212459031

Li, H., Hunter, C. V., & Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*(3), 391-409.

Lim, H. W., & Lee, S. T. (2017). Assessing children's native language in Mandarin using the adapted new Reynell developmental language scales-Mandarin (NRDLS-M). *GEMA Online Journal of Language Studies, 17*(2), 123-145. doi:10.17576/gema-2017-1702-08

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (1996). The Rasch model cannot be "disproved"! *Rasch Measurement Transactions, 10*(3), 512-514.

Linacre, J. M. (2006). Winsteps® (Version 3.66.0) [Computer software]. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2009). Unidimensional models in a multidimensional world. *Rasch Measurement Transactions, 23*(2), 1209.

Linacre, J. M. (2015, June 26). Re: Person fit residuals [Electronic mailing list message]. Retrieved from https://mailman.wu.ac.at/mailman/listinfo/rasch

Linacre, J. M. (2017). *A user's guide to Winsteps Ministep Rasch-model computer programs.* Beaverton, OR: Winsteps.com.

Littlewood, W. (2011). Communicative language teaching: An expanding concept for a changing world. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 541-557). New York, NY: Routledge.

Lock, G. (1996). *Functional English grammar: An introduction for second language teachers.* New York, NY: Cambridge University Press.

Lockwood, J. (2013). The diagnostic English language tracking assessment (DELTA) writing project: A case for post-entry assessment policies and practices in Hong Kong universities. *Papers in Language Testing and Assessment, 2*(1), 30-49.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.

Lynn, P. (2016, August). *The advantage and disadvantage of implicitly stratified sampling.* Swindon, UK: Economic and Social Research Council.

Mackey, A., & Gass, S. M. (2012). Introduction. In A. Mackey, & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 1-4). West Sussex, UK: Blackwell Publishing Ltd.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329-344. doi:10.1177/0265532213480129

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory*. New York, NY: Routledge.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McNamara, T. F. (1996). *Measuring second language performance*. London, UK: Longman.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods.* Advance online publication. doi:10.1037/met0000144

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.* Princeton, NJ: Educational Testing Service.

Meyer, J. P. (2014). *Applied measurement with jMetrik.* New York, NY: Routledge.

Millman, J., & Green, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York, NY: Macmillan.

Ministry of Education Malaysia. (2010, January 14). *Pelaksanaan dasar memartabatkan bahasa Malaysia dan memperkukuh bahasa Inggeris (MBMMBI)* [Implementation of the policy to uphold the Malay language and strengthen the English language]. Putrajaya, Malaysia: Author.

Ministry of Education Malaysia. (2012). *Preliminary report: Malaysia education blueprint 2013-2025.* Putrajaya, Malaysia: Author.

Ministry of Education Malaysia. (2015a). *Kurikulum standard sekolah rendah bahasa Inggeris SK tahun enam: Dokumen standard kurikulum dan pentaksiran* [Primary school standard curriculum for year six national school English language: Standard document for curriculum and assessment]. Putrajaya, Malaysia: Author.

Ministry of Education Malaysia. (2015b). *Kurikulum standard sekolah rendah bahasa Inggeris SJK tahun enam: Dokumen standard kurikulum dan pentaksiran* [Primary school standard curriculum for year six national-type school English language: Standard document for curriculum and assessment]. Putrajaya, Malaysia: Author.

Ministry of Education Malaysia. (2017). *Kurikulum standard sekolah menengah bahasa Inggeris tingkatan 1: Dokumen standard kurikulum dan pentaksiran* [Secondary school standard curriculum for form 1 English language: Standard document for curriculum and assessment]. Putrajaya, Malaysia: Author.

Mizumoto, A., Sasao, Y., & Webb, S. A. (2017). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing,* 1-23. doi:10.1177/0265532217725776

Morrow, K. (1981). Communicative language testing: Revolution or evolution. In J. C. Alderson, & A. Hughes (Eds.), *Issues in language testing* (pp. 9-25). London, UK: British Council.

Mourdoukoutas, P. (2013). Ten leadership quotes from James Cash Penney. Retrieved from https://www.forbes.com/sites/panosmourdoukoutas/2013/02/28/ten-leadership-quotes-from-james-cash-penney/#4371310577ca

Murali, R. S. N. (2015, November 9). 1,000 students drop out due to poor command of the language. The Star Online. Retrieved from http://www.thestar.com.my/

Newton, P. E. (2013, February 4). *Does it matter what validity means?* Archives of the Institute of Education, University of London, England.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment.* Cambridge, UK: Cambridge Assessment.

Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). New York, NY: Kluwer Academic Publishers.

Osterlind, S. J. (2006). *Modern measurement: Theory, principles, and applications of mental appraisal.* Upper Saddle River, NJ: Pearson Education.

Pawlikowska-Smith, G. (2002). *Canadian language benchmarks theoretical framework.* Canada: Centre for Canadian language benchmarks.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Povera, A. (2015, November 25). Making English Sarawak's second official language. *New Straits Times*. Retrieved from http://www.nst.com/

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Purpura, J. E. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 53-68). New York, NY: Springer.

Rabinovich, S. G. (2005). *Measurement errors and uncertainties: Theory and practice* (3rd ed.). New York, NY: Springer.

Rabinovich, S. G. (2013). *Evaluating measurement accuracy: A practical approach* (2nd ed.). New York, NY: Springer.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: The University of Chicago.

Rasch, G. (1964). *An individual-centred approach to item analysis with two categories of answers.* Retrieved from https://www.rasch.org/memo19642.pdf

Rasch, G. (1966). *An individualistic approach to item analysis.* Retrieved from https://www.rasch.org/memo19642.pdf

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Taylor & Francis.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Riget, P. N., & Wang, X. (2016). English for the indigenous people of Sarawak: Focus on the Bidayuhs. In T. Yamaguchi, & D. Deterding (Eds.), *English in Malaysia: Current use and status* (pp. 102-122). Leiden, The Netherlands: Koninklijke Brill.

Ross, K. N. (2005). *Sample design for educational survey research.* Paris, France: UNESCO International Institute for Educational Planning.

Rossiter, J. R. (2011). *Measurement for the social sciences.* New York, NY: Springer.

Rosyidi, A., & Purwati, O. (2017). Revealing intercultural communicative competence in an EFL high school textbook. *Advances in Social Science, Education and Humanities Research, 108,* 65-69.

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foreman and Company.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: Guilford Press.

Samejima, F. (1977). The use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

Sarawak CM to continue Adenan's English policy. (2017, January 23). *The Star Online*. Retrieved from http://www.thestar.com.my/

Sarawak Convention Bureau. (2018). *About Sarawak.* Retrieved from http://sarawakcb.com/sarawak-destination/about-sarawak/

Sarawak Government. (2017). *Sarawak population.* Retrieved from https://www.sarawak.gov.my/web/home/article_view/240/175/

Sarawak State Education Department. (2017). *Data on number of students in form 1 and form 2 according to secondary schools in Sarawak 2017* [Data set]. Sarawak, Malaysia: Author.

Savignon, S. J. (1997). *Communicative competence: Theory and classroom practice* (2nd ed.). New York, NY: Mc-Graw-Hill.

Scheaffer, R. L., Mendenhall, W., III, & Ott, R. L. (2006). *Elementary survey sampling* (6th ed.). Belmont, CA: Thomson Higher Education.

Scheaffer, R. L., Mendenhall, W., III, Ott, R. L., & Gerow, K. (2012). *Elementary survey sampling* (7th ed.). Boston, MA: Brooks/Cole.

Scriven, M. (2002). Assessing six assumptions in assessment. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 255-275). Mahwah, NJ: Lawrence Erlbaum Associates.

Segall D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354. doi:10.1007/BF02294343

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2014). *Measurement theory in action: Case studies and exercises* (2nd ed.). New York, NY: Routledge.

Smyk, E., Restrepo, M. A., Gorin, J. S., & Gray, S. (2013). Development and validation of the Spanish-English language proficiency scale (SELPS). *Language, Speech, and Hearing Services in Schools, 44*, 252-265. doi:10.1044/0161-1461(2013/12-0074)

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 5*, 201-293.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271-295.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

Syed Jaymal Zahiid. (2015, November 11). PM: Poor English eroding Malaysian graduates' self-belief. Malay Mail Online. Retrieved from http://www.themalaymailonline.com/

Taylor, C. S. (2013). *Validity and validation: Understanding statistics.* New York, NY: Oxford University Press.

Ten, M. (2017, July 7). New circular for reps on English proficiency. *Borneo Post Online*. Retrieved from http://www.theborneopost.com/

The University of Auckland. (2016). *DELNA: Diagnostic English language needs assessment (Handbook for candidates at the University of Auckland).* Auckland, New Zealand: Author.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Ed.), *Score reliability* (pp. 43-58). Thousand Oaks, CA: Sage.

Thompson, B., & Vacha-Haase, T. (2003). Psychometrics is datametrics: The test is not reliable. In B. Thomson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 123-148). Thousand Oaks, CA: Sage Publications.

van Bon, W. H. J. (1992). Dimensions in grammatical proficiency. In L. Verhoeven, & J. H. A. L. de Jong (Eds.), *The construct of language proficiency: Applications of psychological models to language assessment* (pp. 33-48). Philadelphia, PA: John Benjamins Publishing Company.

van Steensel, R., Oostdam, R., & van Gelderen, A. (2012). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language Testing, 30*(1), 3-21. doi:10.1177/0265532212440950

Varma, S. (2010). *Preliminary item statistics using point-biserial correlation and p-values.* Morgan Hill, CA: Educational Data Systems.

Wackerly, D. D., Mendenhall, W., III, & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Belmont, CA: Thomson.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149.

We need to be realistic and practical: Sarawakians on using English as an official language. (2015, December 7). *Malaysian Digest*. Retrieved from http://malaysiandigest.com/

What Works Clearinghouse. (2008). *Procedures and standards handbook version 2.* Washington, DC: U.S. Department of Educations, Institute of Education Sciences.

Wheeler, J. A. (1979). From the big bang to the big crunch. *Cosmic Search, 1*(4). Retrieved from http://www.bigear.org/vol1no4/wheeler.htm

Wong, C. W. (2015, June 22). The sad state of English in Malaysia. The Straits Time. Retrieved from http://www.straitstimes.com/

Wright, B. D. (1967, October). *Sample-free test calibration and person measurement*. Paper presented at Educational Testing Service Invitational Conference on Testing Problems, Princeton, NJ.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97-116.

Wright, B. D. (1978, March). *The Rasch model for test construction and person measurement.* Paper presented at the Fifth Annual Conference and Exhibition on Measurement and Evaluation, Los Angeles.

Wright, B. D. (1992). Point-biserials and item fits. *Rasch Measurement Transactions, 5*(4), 174.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*(3), 509-511.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model.* Chicago, IL: The University of Chicago.

Wright, B., & Mead, R. (1979). *Investigating fit with the Rasch model.* Retrieved from https://www.rasch.org/memo74.pdf

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest Manual. Retrieved from https://support.alcatelonetouch.us/hc/en-us/articles/115002315268-Conquest-User-Manual

Wu, M. L., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice.* Gateway East, Singapore: Springer.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne, Australia: Educational Measurement Solutions.

Yamaguchi, T., & Deterding, D. (2016). English in Malaysia: Background, status and use. In T. Yamaguchi, & D. Deterding (Eds.), *English in Malaysia: Current use and status* (pp. 3-22). Leiden, The Netherlands: Koninklijke Brill.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*, 377-392.

Yen, M. W., & Allen, J. M. (1979). *Introduction to measurement theory*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Yi, Y.-S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing, 34*(3), 337-355. doi:10.1177/0265532216646141

Yuen, M. (2015, November 15). Poor English a major handicap. The Star Online. Retrieved from http://www.thestar.com.my/

Zhao, Z. (2013). Diagnosing the English speaking ability of college students in China – Validation of the diagnostic college English speaking test. *RELC Journal, 44*(3), 341-359. doi:10.1177/0033688213500581